CONTRIBUTIONS TO STATISTICAL LEARNING AND STATISTICAL QUANTIFICATION IN NANOMATERIALS

A Thesis
Presented to
The Academic Faculty

by

Xinwei Deng

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2009

CONTRIBUTIONS TO STATISTICAL LEARNING AND STATISTICAL QUANTIFICATION IN NANOMATERIALS

Approved by:

Dr. C. F. Jeff Wu, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Ming Yuan, Co-advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Roshan Joseph Vengazhiyil
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Xiaoming Huo
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Zhonglin Wang
School of Materials Science and Engineering
Georgia Institute of Technology

Date Approved: 18 June 2009

To my parents.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deep gratitude to my advisor, Professor C. F. Jeff Wu. His inspiration, guidance, encouragement and insight have helped and supported me at all phases of my doctoral study. He took care of me both academically and personally in every conceivable way. Not only my academic advisor, he has also been a great mentor for my graduate life.

I am extremely grateful to Dr. Ming Yuan, my co-advisor, for his guidance on my research and active support and encouragement during my studies. His insight and inspiration have greatly influenced the direction and interests of my research. This thesis would not be possible without stimulating discussions from Ming.

I would like to thank Dr. Zhong Lin Wang, Dr. Xiaoming Huo and Dr. Roshan Joseph Vengazhiyil for serving on my dissertation committee and for their valuable comments and suggestions. I would like to acknowledge my sincere debt to Dr. Roshan Joseph Vengazhiyil for his overwhelming support to my research. His constant inspiration has helped me overcome many problems and achieve this milestone. My thanks also go to Dr. Agus Sudjianto at Bank of America for generously supporting me for two summers and inspiring me in various ways.

I am very thankful to my lab members Dr. Tirthankar Dasgupta, Dr. Abhyuday Mandal, Dr. Zhiguang Qian, Dr. Ying Hung, Lulu Kang, Nagesh Adiga, Huizhi Xie, and Matthias Tan who shared time, space and knowledge with me at Georgia Tech. I consider myself very fortunate to be able to spend a lot of time with these outstanding people.

Last, but by no means the least, I would like to give my heartfelt appreciation

and gratitude to my parents and my brother, for their constant support and encouragement during this challenging journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis consists of two parts. The first part focuses on statistical learning and its applications. The second part deals with statistical quantification in nanomaterials.

The first part of this thesis is composed of several work in statistical learning, including methodology, computation and applications. In the first chapter, a new method of Gaussian covariance matrix estimation is developed. The second chapter proposes a new approach to address the robustness of the kernel principal component. An active learning via sequential design, with applications to detection of money laundering, is proposed in the third chapter. Chapter four proposes factor logit-models with a large number of categories. The second part of the thesis is included in chapter five, where we develop a statistical approach to quantifying the elastic deformation of nanomaterials.

The research topic in chapter one is covariance matrix estimation for a large number of Gaussian random variables, which is a challenging yet increasingly common problem. A fact neglected in practice is that the random variables are frequently observed with certain temporal or spatial structures. Such a problem arises naturally in many practical situations with time series and images as the most popular and important examples. Effectively accounting for such structures not only results in more accurate estimation but also leads to models that are more interpretable. In this chapter, we propose shrinkage estimators of the covariance matrix specifically to address this issue. The proposed methods exploit sparsity in the inverse covariance matrix in a systematic fashion so that the estimate conforms with models of Markov

structure and is amenable for subsequent stochastic modeling. The present approach complements the existing work in this direction that deals exclusively with temporal orders and provides a more general and flexible alternative to explore potential Markov properties. It is shown that the estimation procedure can be formulated as a semi-definite program and efficiently computed. The merits of these methods are illustrated through simulation and the analysis of a real data example.

Extending the classical principal component analysis (PCA), the kernel PCA (Schölkopf, Smola and Müller, 1998) effectively extracts nonlinear structures of high dimensional data. As in PCA, the kernel PCA can be sensitive to outliers. Various approaches have been proposed in the literature to robustify the classical PCA. However, it is not immediately clear how these approaches can be "kernelized" in practice. In the second chapter, we propose a robust kernel PCA procedure. We show that the proposed method can be easily computed. Simulations and a real example in the financial service also demonstrate the competitive performance of the proposed approach when there are outlying observations.

The third chapter deals with active learning via sequential design. Motivated by a problem in detecting money laundering accounts, we propose an active learning method using Bayesian sequential designs. The method uses a combination of stochastic approximation and D-optimal designs to judiciously select the accounts for investigation. The sequential nature of the method helps to identify the suspicious accounts with minimal time and effort. An application to real banking data is used to demonstrate the performance of the method. A simulation study shows the efficiency and accuracy of the proposed method, as well as its robustness to model assumptions.

The factor logit-models with a large number of categories are developed in chapter four. We study the theoretical properties of the estimated classifier functions. It is

worth noting that when the number of categories is relatively large, the classifier functions are likely to be located in a functional subspace with much smaller dimensions than the number of categories. Therefore, we propose a factor model for the classifier functions. We show that the convergence rate of the classifier functions estimated from the factor model does not rely on the number of categories, but only on the number of factors. The proposed method therefore can achieve better classification accuracy.

In chapter five, a statistical approach is presented to quantifying the elastic deformation of nanomaterials. Quantifying the mechanical properties of nanomaterials is challenged by its small size, difficulty of manipulation, lack of reliable measurement techniques, and grossly varying measurement conditions and environment. A recently proposed approach is to estimate the elastic modulus from a force-deflection physical model (simply-supported beam model) based on the continuous bridged-deformation of a nanobelt using an Atomic Force Microscope tip under different contact forces. However, the nanobelt may have some initial bending, surface roughness and imperfect physical boundary conditions during measurement, leading to large systematic errors and uncertainty in data quantification. We propose a new statistical modeling technique, called sequential profile adjustment by regression (SPAR), to account for and eliminate the various experimental errors and artifacts. SPAR can automatically detect and remove the systematic errors and therefore gives more precise estimation of the elastic modulus. This work presents an innovative approach that can potentially have a broad impact in quantitative nanomechanics and nanoelectronics.

CHAPTER I

LARGE GAUSSIAN COVARIANCE MATRIX ESTIMATION WITH MARKOV STRUCTURES

1.1 Introduction

In the Gaussian covariance matrix estimation problem, one wishes to estimate the covariance matrix of a multivariate normal vector $X = (X^{(1)}, \ldots, X^{(p)})'$ given an independent and identically distributed sample $X_1, \ldots, X_n$ of X. Assuming that $X \sim \mathcal{N}(\mu, \Sigma)$, μ is typically estimated by the sample mean $\bar{X} = (\bar{X}^{(1)}, \ldots, \bar{X}^{(p)})'$ where

$$\bar{X}^{(i)} = \frac{1}{n} \sum_{j=1}^{n} X_j^{(i)}, \tag{1}$$

and Σ by the sample covariance matrix

$$\widehat{\Sigma}^{\text{SAMPLE}} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)' \left(X_i - \bar{X}\right). \tag{2}$$

Increasingly common in practice, we need to estimate the covariance matrix when the dimension p is moderate or large. It is well known that $\widehat{\Sigma}^{\text{SAMPLE}}$ is not a stable estimate in such cases because of the large number of unknowns involved. Even worse, when $p \geq n$, $\widehat{\Sigma}^{\text{SAMPLE}}$ is not positive definite and therefore not a legitimate covariance matrix estimator for many purposes.

In recent years, a number of new methods have been developed to overcome these drawbacks of the sample covariance matrix. Earlier developments have focused on shrinking the eigenvalues of the sample covariance matrix (Stein, 1977; Haff, 1980; Dey and Srinivasan, 1985; Perron, 1992). Similar idea of perturbing the eigenvalues of the sample covariance matrix also appears in the approach of Ledoit and Wolf (2003) who considered a linear combination of the sample covariance matrix and the

identity matrix. Bayesian treatment of covariance matrix estimation can also be found in Smith and Kohn (2002) and Wong, Carter and Kohn (2003) and references therein. Covariance matrix estimation is closely related to the covariance selection problem (Dempster, 1972) where the interest is in constructing a graphical model that can be used to describe the conditional independence structure among the variables. Yuan and Lin (2007) proposed penalized likelihood methods to simultaneously addressing both problems. Denote $C = (c_{ij}) = \Sigma^{-1}$. A zero entry $c_{ij} = 0$ indicates zero partial correlation between the two random variables $X^{(i)}$ and $X^{(j)}$ and therefore conditional independence given the other variables. The shrinkage estimators of Yuan and Lin (2007) encourage sparsity in the inverse covariance matrix and thus conduct estimation and selection at the same time. Correspondence with a sparse graphical model makes these covariance matrix estimators more interpretable.

A fact neglected by these existing methods is that the random variables are often observed with certain temporal or spatial structures, which arises naturally in the analysis of time series or images. One exception is the approach pioneered by Pourahmadi (1999; 2000) who considered the case when the variables are temporally ordered. Pourahmadi suggested to work on a modified Cholesky decomposition of the covariance matrix: $T\Sigma T' = D$ where

$$T = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \phi_{21} & 1 & 0 & \dots & 0 \\ \phi_{31} & \phi_{32} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \phi_{p1} & \phi_{p2} & \phi_{p3} & \dots & 1 \end{pmatrix}$$

is a lower-triangular matrix with ones on its diagonal and D is a diagonal matrix. It can be shown that the sub-diagonal entries on the ith row of T, $(\phi_{i1}, \dots, \phi_{i,i-1})$ can be interpreted as the minus of the coefficients when regressing $X^{(i)}$ over $X^{(1)}, \dots, X^{(i-1)}$. This provides a natural reparametrization of the covariance matrix when $X^{(i)}$s are

ordered temporally such as in time series. Various shrinkage methods have been proposed within this framework to encourage sparsity in T (Wu and Pourahmadi, 2003; Huang, Liu, Pourahmadi and Liu, 2006; Bickel and Levina, 2007; Levina, Rothman and Zhu, 2007). In particular, Levina et al. (2007) introduced a penalized likelihood estimate that encourages the sparsity of the inverse covariance matrix by forcing a particular pattern of sparsity on T. Note that $\Sigma^{-1} = T'D^{-1}T$. By requiring $\phi_{ij} = 0$ if $\phi_{i,j+1} = 0$ and $j < i - 1$, some entries of the inverse covariance matrix that are far away from the diagonal can be shrunken to zeros and therefore the estimate can be interpreted as Markov chains. These approaches, however, only apply to temporal orders and may not be suitable if the $X^{(i)}$s are observed with more complicated structures such as spatial orders.

To elaborate, consider analyzing handwritten digits based on a training sample of images (LeCun et al., 1990) as shown in Figure 1. Covariance matrix estimation of the

Figure 1: Sample images of handwritten digits: each image is of size 16×16.

intensity values on the $256 = 16 \times 16$ pixels plays a critical role in various statistical analysis such as principal component analysis and linear discriminant analysis. The

correlation between the intensity on two pixels is clearly related to the positions of the pixels. Furthermore, images of this sort can most often be adequately modeled as a Markov random field of a relatively small order since the intensity values on pixels far away from each other generally are independent of each other conditional on intensities of the other pixels (Winkler, 2006). A covariance matrix estimate that conforms with such models not only reduces the dimensionality of the estimation problem but also is much more valuable in subsequent stochastic modeling. Unlike the Markov structure in the temporally ordered cases, the Markov random field can not be inferred from the sparsity pattern of matrix T of the modified Cholesky decomposition. To illustrate, consider four random variables that are observed from a 2×2 grid as shown below.

$X^{(1)}$	$X^{(2)}$
$X^{(3)}$	$X^{(4)}$

A Markov random field of order one is equivalent to $c_{41} = c_{32} = 0$, which can only imply that $\phi_{41} = 0$ as illustrated by (3).

$$\Sigma^{-1} = \begin{pmatrix} 1 & 0.4 & 0.4 & 0 \\ 0.4 & 1 & 0 & 0.4 \\ 0.4 & 0 & 1 & 0.4 \\ 0 & 0.4 & 0.4 & 1 \end{pmatrix} \Longrightarrow T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.59 & 1 & 0 & 0 \\ 0.48 & -0.19 & 1 & 0 \\ 0 & 0.4 & 0.4 & 1 \end{pmatrix} \tag{3}$$

But on the other hand, a Markov random field of order one can not be inferred from $\phi_{41} = 0$. A counterexample is given by (4). This simple example shows that the modified Cholesky decomposition may no longer be suitable for exploring Markov structures when the variables are observed with structures more general than temporal orders.

$$\Sigma^{-1} = \begin{pmatrix} 1 & 0.28 & 0.4 & 0 \\ 0.28 & 1.18 & -0.26 & 0.4 \\ 0.4 & -0.26 & 1 & 0.4 \\ 0 & 0.4 & 0.4 & 1 \end{pmatrix} \Longleftarrow T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.73 & 1 & 0 & 0 \\ 0.48 & -0.5 & 1 & 0 \\ 0 & 0.44 & 0.44 & 1 \end{pmatrix} \tag{4}$$

The lack of a method that can handle general Markov structures among the random variables motivates the present work. In this chapter, we propose a more direct strategy to explore conditional independence relationships among variables when they are observed with temporal and spatial structures. We suggest to exploit sparsity directly on the inverse covariance matrix. We consider constrained maximum likelihood methods with constraints that encourage sparsity in a systematic fashion so that estimates that conform with models of Markov structure are favored. We shall introduce the proposed methods in the next section, followed by examples in Sections 1.3 and 1.4. We conclude with some discussions in the last section.

1.2 Methodology

The log likelihood for μ and $C = \Sigma^{-1}$ based on a random sample $X_1, \ldots, X_n$ of X is

$$\ln |C| - \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)' C (X_i - \mu) = \ln |C| - \text{trace}(C\bar{A}) \tag{5}$$

up to a constant not depending on μ and C, where

$$\bar{A} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right) \left(X_i - \bar{X}\right)' \tag{6}$$

is the maximum likelihood estimate of Σ. To estimate C, we consider a shrunken version of $\tilde{C} = \bar{A}^{-1}$: $C = (\theta_{ij}\tilde{c}_{ij})$ where θ'_{ij}s are shrinkage coefficients. Other choices of $\tilde{C}$ are also possible; we focus on $\bar{A}^{-1}$ in this chapter to fix ideas. Given that $\tilde{C}$ is a reasonably good initial estimate of the inverse covariance matrix it is appropriate to require that the shrinkage coefficients be nonnegative, $\theta_{ij} \geq 0$. To achieve sparse graph structure and encourage sparsity in C, one can maximize the log likelihood subject to the constraint that

$$\sum_{i \neq j} \theta_{ij} \leq M \tag{7}$$

for some tuning parameter $M \geq 0$. This is the so-called graphGarrote estimator proposed by Yuan and Lin (2007). Clearly when $M = +\infty$, the constraint becomes

inactive and the resulting estimate reduces to $\bar{A}^{-1}$ and no shrinkage takes place. On the other hand when $M = 0$, all the off-diagonal entries of the inverse covariance matrix will be shrunken to zero and the estimate becomes diagonal which implies mutual independence among $X^{(i)}$s. A choice of tuning parameter M between these two extremes will result in covariance matrix estimates with varying degrees of sparsity. The procedure is similar in spirit to the nonnegative garrote estimator proposed by Breiman (1995) for linear regression.

We now consider the situation when the random variables are observed in a space with a certain distance measure defined. Assume that $X^{(i)}$ is observed at location t_i. For example, t_i is a point in a two dimensional lattice in the case of images. Most often dependence between two variables dwindles as the distance between them increases. To incorporate this prior information into the estimation of the covariance matrix, we impose the following constraints on the shrinkage coefficients.

$$\theta_{ij} \le \theta_{ik} \quad \text{if} \quad d_{ij} \ge d_{ik} \tag{8}$$

where $d_{ij} = \text{dist}(t_i, t_j)$ is the pairwise distance. Because the entries of $\tilde{C}$ are generally nonzero, constraint (8) implies that $c_{ij} = 0$ if $c_{ik} = 0$. It is worth pointing out that this constraint only encourages more shrinkage towards 0 for entries that are farther away from the diagonal to reflect our preference towards Markov models; it does not force $c_{ij} \le c_{ik}$. In summary, we propose to estimate C by $\widehat{C} = (\widehat{\theta}_{ij}\tilde{c}_{ij})$ where $\widehat{\Theta} = (\widehat{\theta}_{ij})$ is the solution to

$$\begin{aligned} \min \; & -\left[\ln|C| - \text{trace}(C\bar{A})\right] \\ \text{subject to} \quad & C \text{ is positive definite} \\ & c_{ij} = \theta_{ij}\tilde{c}_{ij} \\ & \theta_{ij} \ge 0 \\ & \sum_{i \ne j} \theta_{ij} \le M \\ & \theta_{ij} \le \theta_{ik} \quad \text{if} \quad d_{ij} \ge d_{ik} \text{ and } j, k \ne i. \end{aligned} \tag{9}$$

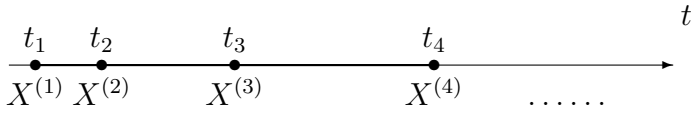
The problem is a semi-definite program and can be easily solved using standard software packages such as `SDPT3` (Tütüncü, Toh and Todd, 2003).

Thus far we have assumed that the tuning parameter M is fixed. In practice, it also needs to be estimated. A commonly used approach is the multi-fold cross-validation which can be computationally demanding. A much more efficient alternative is the BIC criterion introduced by Yuan and Lin (2007):

$$\mathrm{BIC}(M) = -\ln|\hat{C}(M)| + \mathrm{trace}(\hat{C}(M)\bar{A}) + \frac{\ln(n)}{n}\sum_{i\le j}\hat{e}_{ij}(M), \tag{10}$$

where $\hat{e}_{ij} = 0$ if $\hat{c}_{ij} = 0$, and $\hat{e}_{ij} = 1$ otherwise. We shall adopt this criterion in our implementation and it works very well in practice according to our experience.

Note that without the last constraint in (9), our estimate becomes the graph-Garrote of Yuan and Lin (2007). The last constraint takes the temporal or spatial structure of the observation into consideration. Consider, for example, the case where the observations are temporally ordered.



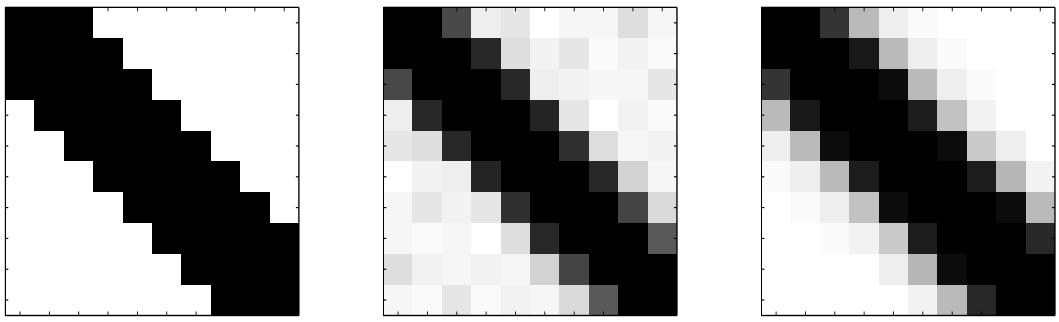
Because d_{ij} is a monotone increasing transformation of $|i-j|$, the last constraint can be simplified to

$$\theta_{i,j+1} \le \theta_{ij} \quad \text{if } j > i, \text{ and } \theta_{i,j+1} \ge \theta_{ij} \quad \text{if } j < i-1. \tag{11}$$

This encourages more shrinkage to the partial correlation between $X^{(i)}$ and $X^{(j)}$ if the two observations are farther away from each other. Together with the constraint on the sum of the shrinkage coefficients, it induces a sparse estimate of the inverse covariance matrix that follows a non-stationary Markov chain in that there exist $h_1, h_2, \ldots, h_p > 0$ such that

$$X^{(i)} \perp \{X^{(j)} : d_{ij} > h_i\} \big| \{X^{(j)} : 0 < d_{ij} \le h_i\}.$$

To demonstrate its effect, we apply both graphGarrote and the proposed estimate, hereafter we refer to as the structured graphGarrote, to data sets that are simulated from a AR(2) model with $c_{ij} = 1$ if $i = j$, 0.5 if $|i - j| = 1$, 0.25 if $|i - j| = 2$ and 0 otherwise. We consider sample size $n = 100$ and dimension $p = 10$. For both estimates, the tuning parameter M is chosen by the BIC criterion defined by (10). Panel (a) of Figure 2 shows the nonzero pattern of the true inverse covariance matrix. A black block indicates that the coefficient is not zero and a white block corresponds to a zero entry. Panels (b) and (c) give the heatmap representing the frequency that each entry of the inverse covariance is estimated as nonzero over 100 simulations. A darker block indicates higher frequency. It is evident that by taking advantage of the temporal order, the proposed method is more suitable to exploit the Markov structure of the true data generating mechanism.



(a) True (b) graphGarrote (c) Structured graphGarrote

Figure 2: Effect of the structure constraint: Panel (a) represents the true nonzero pattern of the inverse covariance matrix with a block at the ith row and jth column indicating $c_{ij} \neq 0$; Panels (b) and (c) give the frequency that each entry of the inverse covariance matrix is estimated by a nonzero value. A darker block indicates higher frequency. Panel (b) correspond to graphGarrote and (c) corresponds to the structured graphGarrote.

In the case of images, the random variables are observed on a two dimensional lattice. Let $X^{(i)}$ be observed at the pixel located on the r_ith row and c_ith column. A natural distance defined on a two dimensional lattice is the so-called city block

distance or Manhattan distance:

$$\text{dist}(i,j) = |r_i - r_j| + |c_i - c_j|. \tag{12}$$

For example, consider $p = 3 \times 3$ random variables that are observed with the following spatial structure. The city block distance between $X^{(1)}$ and other random variables

$X^{(1)}$	$X^{(2)}$	$X^{(3)}$
$X^{(4)}$	$X^{(5)}$	$X^{(6)}$
$X^{(7)}$	$X^{(8)}$	$X^{(9)}$

are given as

$$\begin{aligned} d_{12} = d_{14} &= 1 \\ d_{13} = d_{15} = d_{17} &= 2 \\ d_{16} = d_{18} &= 3 \\ d_{19} &= 4. \end{aligned}$$

It is noteworthy that in this example and many others as well, the pairwise distances between observations take only a few distinct values. It is natural to expect that similar degrees of shrinkage is needed for entries of the inverse covariance matrix that correspond to similar pairwise distances. In other words, it is reasonable to have $\theta_{ij} = \theta_{i'j'}$ if $d_{ij} = d_{i'j'}$. In the current example, this amounts to

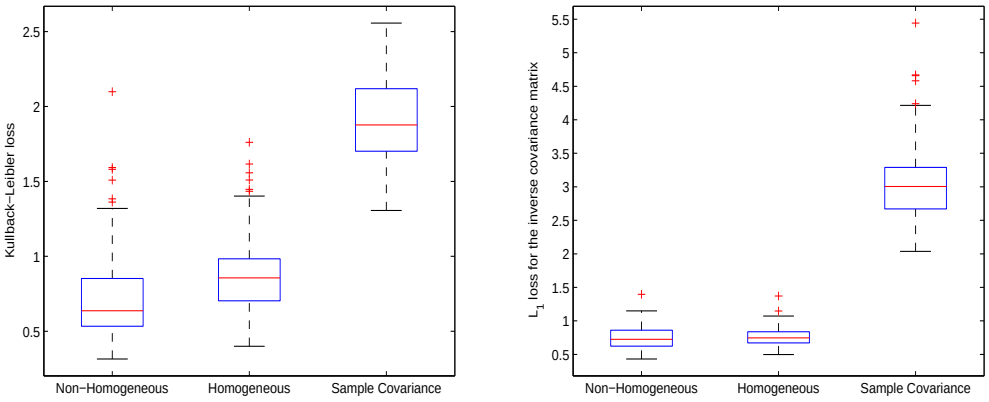
$$\begin{aligned} \theta_{12} = \theta_{14} &\equiv \theta_1, \\ \theta_{13} = \theta_{15} = \theta_{17} &\equiv \theta_2, \\ \theta_{16} = \theta_{18} &\equiv \theta_3, \\ \theta_{19} &\equiv \theta_4. \end{aligned}$$

For convenience, we shall refer to this modification as the homogeneous structured graphGarrote estimate. It is worth pointing out that the homogeneous estimate does not impose stationarity by forcing $c_{ij} = c_{i'j'}$. It, however, greatly reduces the dimensionality of the optimization problem (9), which brings about great computational

efficiency. The difference in estimation accuracy between the structured graphGarrote and its homogeneous version is generally marginal. To illustrate this, consider $p = 4 \times 4$ random variables that are observed on a two dimensional lattice. We generate $n = 100$ observations from a multivariate normal distribution with the inverse covariance matrix generated in the following fashion. First we generate $c_{ij} \sim U(0, 1)$ if $d_{ij} = 1$, and set $c_{ij} = 1$ if $d_{ij} = 0$ and 0 if $d_{ij} > 1$. Here d_{ij} represents the city block distance between i and j, and $U(0, 1)$ denotes the uniform distribution from 0 to 1. Next for all i, we normalize c_{ij} so that $\sum_{i \neq j} c_{ij} = 0.9$ to ensure positive definiteness. We apply both the structured graphGarrote and its homogeneous version to the simulated data. We also include the sample covariance matrix in the comparison to serve as the baseline. Figure 3 shows the boxplot of the estimation accuracy measured by both the Kullback Leiber loss and the matrix ℓ_1 loss for the three methods, summarized over 100 simulated data sets. Both criteria will be defined in the next section. We observe from Figure 3 that even if the true data generating mechanism is non-stationary as in this example, the structured graphGarrote and its homogeneous version behave very similarly. Because of the similarity in estimation accuracy and the great computational advantage of the homogeneous version, we shall use it throughout the chapter unless otherwise indicated.

1.3 Simulations

In this section, we compare through simulations the proposed methods with several popular alternative shrinkage estimators of the covariance matrix including those of Bickel and Levina (2006), Huang et al. (2006), Levina et al. (2007) as well as the sample covariance matrix. We compare these methods on the basis of the number of false positives (FP; incorrectly identified nonzero entries of Σ^{-1}), the number of false negatives (FN; incorrectly missed nonzero entries), and several commonly used



(a) Kullback-Leibler loss (b) L_1 loss for the inverse covariance matrix

Figure 3: Comparison of estimation accuracy between the structured graphGarrote and its homogeneous versions. Panels (a) and (b) correspond to loss functions given by (13) and (15) respectively.

estimation accuracy measures, namely the Kullback-Leibler loss defined as

$$\mathrm{KL} = -\log|\hat{C}| + \mathrm{tr}(\hat{C}\Sigma) - (-\log|\Sigma^{-1}| + p), \tag{13}$$

the quadratic loss

$$\mathrm{QL} = \mathrm{tr}(\Sigma^{-1}\hat{\Sigma} - I)^2, \tag{14}$$

where $\hat{\Sigma} = \hat{C}^{-1}$, and the matrix ℓ_1 loss

$$L_1 = \|\Sigma - \hat{\Sigma}\|_{\ell_1}, \tag{15}$$

where $\|M\|_{\ell_1} = \sup\{\|Mx\|_{\ell_1} : \|x\|_{\ell_1} = 1\}$ and $\|x\|_{\ell_1}$ is ℓ_1 norm of vector x.

In the approach of Bickel and Levina (2006), the Cholesky factor T is banded to estimate the inverse covariance matrix, i.e.,

$$\phi_{ij} = 0, \qquad \forall |i-j| > h$$

for some $h > 0$. The banding parameter h is chosen by cross validation using the matrix ℓ_1 loss. Huang et al. (2006) suggested adding ℓ_1 or ℓ_2 penalty

$$\lambda \sum_{i=1}^{p} \sum_{j=1}^{i-1} |\phi_{ij}|^{\gamma}, \qquad \gamma = 1 \text{ or } 2$$

on the elements of T to the normal likelihood (5), which leads to Lasso or ridge type shrinkage of the ϕ_{ij}s. The authors also suggested to choose the tuning parameter $\lambda > 0$ by cross validation. The ℓ_2 penalty ($\gamma = 2$) was used in our simulation study. Instead of ℓ_1 penalty in Huang et al. (2006), Levina et al. (2007) introduced a nested Lasso penalty on ϕ_{ij}s. The so-called J_2 nested Lasso penalty is given by $\sum_j J_{2j}$ where

$$J_{2j} = \lambda_1 \sum_{k=1}^{j-1} |\phi_{jk}| + \lambda_2 \sum_{k=1}^{j-2} \frac{|\phi_{jk}|}{|\phi_{j,k+1}|}, \tag{16}$$

and $\lambda_1, \lambda_2 > 0$ are tuning parameters. The nested Lasso penalty forces a random variable to be conditionally dependent only on its nearest neighbors. Different from banding, the number of nearest neighbors selected with the nested Lasso penalty is allowed to vary across variables. As suggested by the authors, the tuning parameters are selected with a validation set which is set aside from the original training data set.

1.3.1 Temporal Structures

The first set of simulation concerns temporally ordered observations. The following three models were considered.

Model 1. $\Sigma = I_p$.

Model 2. (AR(1) model) $c_{ij} = 1$ if $i = j$, 0.45 if $|i - j| = 1$ and 0 otherwise.

Model 3. (AR(2) model) $c_{ij} = 1$ if $i = j$, 0.5 if $|i - j| = 1$, 0.25 if $|i - j| = 2$ and 0 otherwise.

For each model, we simulated data sets with sample size $n = 100$ and dimension $p = 10$, $n = 100$ and $p = 30$, or $n = 400$ and $p = 100$. Table 1 documents the means and standard errors (in parentheses), summarized from 100 runs for each combination.

As shown in Table 1, all shrinkage methods improve upon the sample covariance matrix. The improvement is particularly significant for high dimensional problems.

Among the shrinkage methods, the structured graphGarrote enjoys the best performance overall in terms of estimation accuracy. It also dominates the other methods overwhelmingly in recovering the nonzero patterns of the inverse covariance matrix or equivalently the Markov structure among the variables.

We have also conducted simulation on a couple of other models considered by Huang et al. (2006) and Bickel and Levina (2006) respectively, which are

Model 4. Covariance matrix such that $\sigma_{ij} = \rho^{|i-j|},\ 1 \le i, j \le p$ with $\rho = 0.5$.

Model 5. $C = T' D^{-1} T$, where $D = 0.01 \times I$, and $T = -(\phi_{ij})$, with $\phi_{ii} = 1$, $\phi_{i+1,i} = 0.8$, and $\phi_{ij} = 0$ otherwise.

Both are AR(1) and the results are very similar to those of Model 1 and therefore omitted here.

When the observations follow Markov chains of varying lengths from variable to variable, the number of nonzero elements differs among the rows of the Cholesky factor matrix T, i.e.,

$$\phi_{ij} = 0 \text{ if and only if } i - j > h_i$$

for different bandwidth h_is. The banding method in Bickel and Levina (2006) may no longer be appropriate since it assumes that $h_1 = \ldots = h_p$. Levina et al. (2007) addressed this by allowing different bandwidths for different rows of T. The non-homogeneous structured graphGarrote can also overcome this problem. To illustrate, consider a model similar to that of Levina et al. (2007).

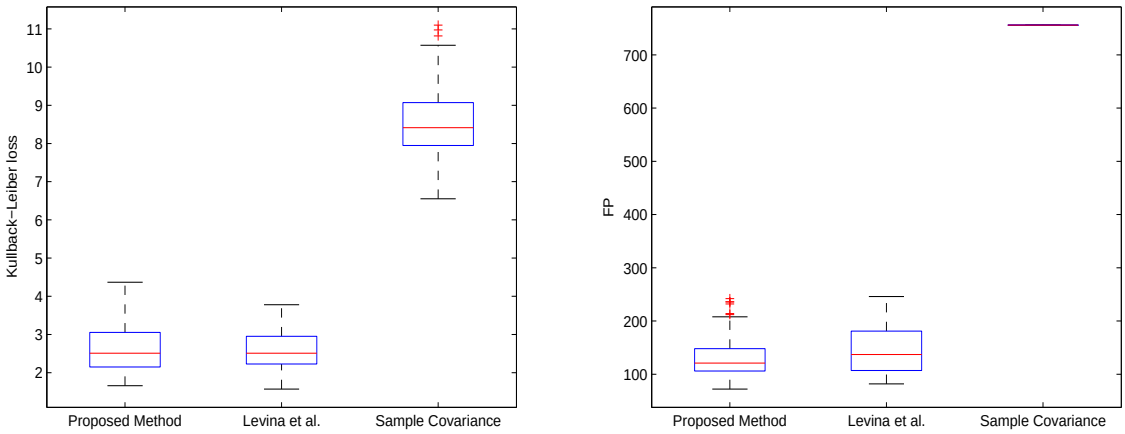
Model 6. $C = (I - \Phi)' D^{-1} (I - \Phi)$, with $D = 0.01 \times I$ and $\Phi = (\phi_{i,j})$ where $\forall j \ge 2$, $k_j \sim U(\lceil j/2 \rceil, j-1)$; $\phi_{j,j'} = 0.5$, $k_j \le j' \le j-1$; $\phi_{i,j} = 0$, $j' < k_j$.

Here $U(k_1, k_2)$ denotes an integer selected randomly from integer k_1 to k_2. In this simulation we take $p = 30$ as an example. To avoid poorly conditioned covariance matrix, we divided the 30 variables into two independent blocks with 15 variables each, and

Table 1: Simulation results for the three models with temporal orders. Averages and standard errors are calculated from 100 runs.

		Structured graphGarrote					Bickel and Levina					Levina et al.					Huang et al.			Sample Covariance		
p	Model	KL	QL	L_1	FP	FN	KL	QL	L_1	FP	FN	KL	QL	L_1	FP	FN	KL	QL	L_1	KL	QL	L_1
	1	0.10	0.19	0.25	0.00	0.00	0.11	0.21	0.27	0.18	0.00	0.11	0.22	0.28	1.84	0	0.12	0.21	0.33	0.69	1.10	1.14
		(0.05)	(0.08)	(0.07)	(0.00)	(0.00)	(0.05)	(0.10)	(0.09)	(1.79)	(0.00)	(0.05)	(0.10)	(0.09)	(5.39)	(0.00)	(0.06)	(0.09)	(0.08)	(0.17)	(0.18)	(0.14)
10	2	0.26	0.63	2.26	0.16	0.00	0.63	1.04	5.87	64.98	0.00	0.26	0.56	2.37	0.72	0	0.60	1.00	6.96	0.68	1.10	6.26
		(0.10)	(0.27)	(0.78)	(1.59)	(0.00)	(0.17)	(0.22)	(2.35)	(8.28)	(0.00)	(0.10)	(0.31)	(1.21)	(7.16)	(0.00)	(0.13)	(0.22)	(2.61)	(0.14)	(0.23)	(2.34)
	3	0.36	0.83	1.66	2.76	0.00	0.71	1.66	2.34	6.90	7.04	0.66	1.13	2.07	53.76	0.64	0.62	1.03	1.95	0.70	1.09	2.06
		(0.11)	(0.33)	(0.32)	(6.06)	(0.00)	(0.38)	(1.15)	(0.72)	(7.95)	(7.94)	(0.21)	(0.54)	(0.58)	(10.97)	(3.13)	(0.16)	(0.21)	(0.30)	(0.18)	(0.21)	(0.40)
	1	0.31	0.58	0.32	0.00	0.00	0.32	0.62	0.33	0.58	0.00	0.32	0.60	0.34	45.52	0.00	0.34	0.59	0.41	8.37	9.15	3.24
		(0.09)	(0.16)	(0.06)	(0.00)	(0.00)	(0.11)	(0.21)	(0.08)	(5.77)	(0.00)	(0.08)	(0.16)	(0.07)	(57.72)	(0.00)	(0.10)	(0.15)	(0.07)	(0.90)	(0.72)	(0.26)
30	2	1.03	2.61	3.88	0.06	0.00	7.26	8.72	52.63	734.60	0.00	0.88	1.86	4.60	125.76	0.00	5.46	7.42	100.90	8.58	9.34	52.82
		(0.23)	(0.92)	(0.68)	(0.60)	(0.00)	(1.44)	(0.96)	(20.76)	(83.69)	(0.00)	(0.17)	(0.42)	(1.69)	(75.16)	(0.00)	(0.63)	(0.81)	(23.14)	(0.94)	(0.80)	(17.97)
	3	1.43	3.54	2.54	0.54	0.00	2.05	4.76	3.04	29.54	15.68	4.59	21.39	7.54	168.1	40.86	4.92	10.92	4.36	8.76	9.16	6.08
		(0.29)	(1.07)	(0.33)	(5.37)	(0.00)	(1.28)	(4.01)	(0.74)	(31.48)	(25.14)	(0.26)	(4.71)	(1.44)	(82.18)	(6.71)	(0.36)	(1.23)	(0.27)	(0.92)	(0.74)	(0.60)
	1	0.26	0.51	0.20	0.00	0.00	0.26	0.52	0.20	5.94	0.00	0.25	0.50	0.20	147.48	0.00	0.35	0.64	0.48	20.14	25.27	4.91
		(0.04)	(0.07)	(0.03)	(0.00)	(0.00)	(0.06)	(0.12)	(0.04)	(33.78)	(0.00)	(0.04)	(0.07)	(0.04)	(103.8)	(0)	(0.05)	(0.07)	(0.04)	(0.74)	(0.62)	(0.17)
100	2	0.81	1.74	2.32	0.00	0.00	18.38	24.24	285.77	9227.58	0.00	1.13	2.25	5.73	6930.58	0.00	13.27	18.90	903.92	20.11	25.29	284.53
		(0.10)	(0.23)	(0.36)	(0.00)	(0.00)	(1.37)	(1.04)	(106.63)	(393.22)	(0.00)	(0.16)	(0.31)	(1.13)	(731.462)	(0)	(0.35)	(0.47)	(129.05)	(0.65)	(0.56)	(106.30)
	3	1.26	2.73	1.66	0.00	0.00	1.31	2.53	1.71	400.20	0.00	8.08	29.28	6.57	7501.46	16.8	12.81	24.14	6.62	20.17	25.10	9.21
		(0.14)	(0.34)	(0.20)	(0.00)	(0.00)	(0.13)	(0.24)	(0.20)	(73.76)	(0.00)	(0.51)	(2.78)	(0.53)	(511.37)	(7.90)	(0.31)	(0.77)	(0.23)	(0.67)	(0.59)	(0.38)

generated a random structure from Model 6 for each block, i.e., $(X^{(1)}, \ldots, X^{(15)})$ and $(X^{(16)}, \ldots, X^{(30)})$ each follows Model 6 but are mutually independent. We simulated samples with size $n = 100$ and compared the structured graphGarrote, the method used in Levina et al. (2007), and the sample covariance matrix. Figure 4 reports the boxplot for the Kullback-Leibler loss (KL) and the number of false positive of the inverse covariance matrix (FP) from 100 runs.



(a) Kullback-Leibler loss (KL)

(b) Number of false positives (FP)

Figure 4: Estimation comparison for Model 6.

From Figure 4, we observe that the performance of our proposed method is very similar to that of Levina et al. (2007). By capturing the structure of the true model, the structured graphGarrote committed relatively fewer false positives. We note that the optimization problem involved in the approach of Levina et al. (2007) is not convex. An iterative procedure was developed by Levina et al. (2007) to tackle the computational challenge. Although efficient, it can still be sensitive to the choice of tuning parameters and initial values. In contrast, the proposed method is strictly convex and more stable in computation.

1.3.2 Spatial Structures

Next, we consider the situation when the random variables are observed on a two dimensional lattice. Three different models were used in our simulation.

Model 7 $\Sigma = I_p$.

Model 8 (Markov random field of order one) $c_{i,j} = 1$ if $d_{ij} = 0$, 0.25 if $d_{ij} = 1$, and 0 otherwise.

Model 9 (Markov random field of order two) $c_{i,j} = 1$ if $d_{ij} = 0$, 0.4 if $d_{ij} = 1$, 0.15 if $d_{ij} = 2$, and 0 otherwise.

For each model, we simulated samples of size $n = 100$ and dimension $p = 4 \times 4$, $n = 200$ and $p = 8 \times 8$, or $n = 600$ and $p = 16 \times 16$. Although in principle, all the methods described previously can be applied in these settings, none of the methods except for the structured graphGarrote is devised to take advantage of the spatial structure explicitly. Table 2 shows that being able to account for the spatial structures, the structure graphGarrote enjoys considerably improved performance over the other methods. The tremendous reduction in estimation error over the sample covariance matrix for large p is particularly noteworthy.

1.4 Handwritten Digit Data

To further illustrate the merits of the proposed method, we apply the proposed structured covariance matrix estimation to a real data example. The handwritten digit data (LeCun et al., 1990) come from automatic reading of handwritten zip codes appeared on envelopes by the United States Postal Service. Each handwritten digit is converted into a 16 by 16 grayscale image after some processing. The intensity values lie in the range from -1 and 1. Images as such can often be modeled as a Markov random field of a relatively small order. The proposed methods exploit the sparsity

Table 2: Simulation results for the three models with spatial structure. Averages and standard errors are calculated from 100 runs.

		Structured graphGarrote					Bickel and Levina					Levina et al.					Huang et al.			Sample Covariance		
p	Model	KL	QL	L_1	FP	FN	KL	QL	L_1	FP	FN	KL	QL	L_1	FP	FN	KL	QL	L_1	KL	QL	L_1
	7	0.17	0.31	0.28	0.00	0.00	0.18	0.34	0.30	0.00	0.00	0.18	0.36	0.31	3.92	0.00	0.18	0.32	0.34	1.92	2.71	1.80
		(0.06)	(0.10)	(0.07)	(0.00)	(0.00)	(0.07)	(0.12)	(0.07)	(0.00)	(0.00)	(0.07)	(0.13)	(0.09)	(8.83)	(0.00)	(0.07)	(0.11)	(0.08)	(0.36)	(0.39)	(0.22)
16	8	0.55	1.22	2.51	0.02	0.00	0.84	1.58	2.32	57.56	1.2	1.75	3.73	3.42	107.86	10.8	1.15	1.84	3.21	1.97	2.75	2.77
		(0.14)	(0.36)	(0.62)	(0.20)	(0.00)	(0.25)	(0.82)	(0.70)	(10.84)	(5.23)	(0.21)	(1.32)	(1.04)	(91.19)	(11.74)	(0.18)	(0.25)	(0.51)	(0.37)	(0.40)	(0.52)
	9	0.96	2.53	3.27	1.9	0.02	2.01	6.69	4.26	20.52	50.34	1.81	2.68	3.34	123.92	0.00	1.52	2.57	3.12	1.86	2.66	3.34
		(0.23)	(0.91)	(0.59)	(10.80)	(0.30)	(0.69)	(4.89)	(0.49)	(8.72)	(25.70)	(0.27)	(0.33)	(0.50)	(0.48)	(0.00)	(0.24)	(0.48)	(0.37)	(0.30)	(0.35)	(0.55)
	7	0.34	0.65	0.26	0.00	0.00	0.35	0.67	0.27	10.04	0.00	0.33	0.65	0.28	108	0.00	0.42	0.75	0.51	19.47	20.79	4.62
		(0.06)	(0.11)	(0.05)	(0.00)	(0.00)	(0.12)	(0.21)	(0.06)	(34.05)	(0.00)	(0.06)	(0.12)	(0.05)	(208.33)	(0.00)	(0.08)	(0.12)	(0.05)	(1.02)	(0.88)	(0.26)
64	8	1.76	4.22	8.64	0.08	0.00	3.05	5.55	6.46	727.32	0.00	10.03	30.60	19.84	1819.64	45.9	7.22	10.06	14.03	19.45	20.71	9.18
		(0.24)	(0.77)	(1.96)	(0.80)	(0.00)	(0.23)	(0.37)	(1.28)	(1.14)	(0.00)	(0.25)	(2.24)	(0.39)	(454.77)	(12.72)	(0.40)	(0.39)	(0.92)	(1.08)	(0.81)	(1.15)
	9	4.37	12.24	10.65	0.00	0.12	9.18	28.16	13.88	522.18	200.34	17.53	97.60	15.59	2208.28	123.2	11.04	25.22	10.55	19.44	20.64	10.82
		(0.53)	(2.20)	(1.07)	(0.00)	(0.47)	(0.82)	(13.59)	(0.80)	(63.00)	(35.82)	(0.45)	(31.86)	(2.21)	(592.24)	(60.87)	(0.47)	(1.92)	(0.70)	(1.01)	(0.79)	(0.86)
	7	0.64	1.26	0.21	0.00	0.00	0.43	0.86	0.18	0.00	0.00	0.84	1.67	0.31	562.16	0.00	1.60	2.78	1.25	355.12	163.95	12.14
		(0.06)	(0.10)	(0.03)	(0.00)	(0.00)	(0.04)	(0.08)	(0.02)	(0.00)	(0.00)	(0.05)	(0.09)	(0.03)	(14.53)	(0.00)	(0.08)	(0.11)	(0.05)	(6.75)	(2.06)	(0.34)
256	8	3.14	7.32	23.98	0.00	0.00	7.55	14.32	16.49	6946.26	0.00	43.18	318.80	73.43	1267.74	404.12	35.63	45.39	52.34	353.68	163.97	30.26
		(0.18)	(0.56)	(4.59)	(0.00)	(0.00)	(0.18)	(0.34)	(2.95)	(5.70)	(0.00)	(0.53)	(16.14)	(1.13)	(196.08)	(11.31)	(0.56)	(0.58)	(2.41)	(7.79)	(1.82)	(3.07)
	9	11.34	29.00	19.96	0.1	0.12	40.31	137.92	28.86	6046.64	898.6	58.83	525.89	25.87	793.32	1274.32	55.99	147.99	23.28	353.16	164.23	31.54
		(0.75)	(2.66)	(1.41)	(0.99)	(0.47)	(0.20)	(2.46)	(0.76)	(5.37)	(1.15)	(0.28)	(12.63)	(0.24)	(107.07)	(31.87)	(0.58)	(1.63)	(0.64)	(8.49)	(1.75)	(1.10)

in the inverse covariance matrix so that the estimate conforms with models of such Markov structure.

A common goal of analyzing the handwritten digits is to distinguish images representing different digits. To this end, we consider applying linear discriminant analysis (LDA) with covariance matrix estimated using the proposed method as well as the sample covariance matrix to the data. The main purpose of this exercise is to demonstrate how the structured graphGarrote can lead to improved classification performance of LDA. For illustrative purpose, we focus on digits 6 and 9 which include a total of 1308 images in the data set. 600 images were randomly selected as the training set, and the rest were used as the test set. We repeated the experiment 100 times. The boxplot of the testing error is given in Figure 5. It shows that the covariance matrix estimated from the proposed method indeed leads to lower misclassification error.

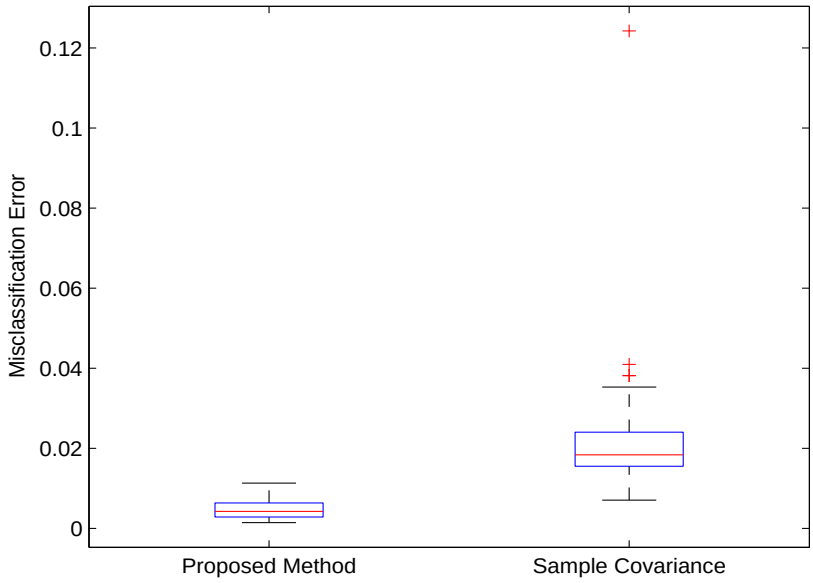


Figure 5: The boxplot of misclassification error on the test set for 100 replications.

To gain further insights, we also examine for a given pixel, how often its partial correlation with other pixels is estimated by a nonzero value. The (i, j) panel of

Figure 6 corresponds to the partial correlation between the intensity at the (i, j)th pixel and other pixels. A darker cell indicating higher frequency. A few pixels around the four corners are removed from our analysis because their intensity values remain constant in the data set. A more detailed look at several selected pixels is given in Figure 7. From Figures 6 and 7, we observe that the handwritten digit images may be modeled by a Markov random field of order 4 or 5.

1.5 Discussions

In this chapter, we have developed methods for estimating high dimensional Gaussian covariance matrix when the random variables are observed with temporal or spatial structures. By directly exploiting sparsity of the inverse covariance matrix, the estimate obeys certain Markov models. The proposed method can be formulated as a semi-definite program and efficiently computed using standard software.

Although we focused on the temporal and spatial structures, the method can be easily extended to more complicated situations such as spatial-temporal structures. More generally, our method can be applied in situations where a similarity/dissimilarity measure of the domain from which the variables are observed is available.

The proposed estimates of the inverse covariance matrix are shrunken version of $\bar{A}^{-1}$. As we pointed out earlier, other initial estimate of Σ can also be employed. For example, we can consider a linear combination of the sample covariance and the identity matrix. This is particularly appealing when $p \geq n$ and the inverse of $\bar{A}$ does not exist. Another initial estimate that might be of great interest in this case is the MLE of C with $c_{ij} = 0$ for $|i - j| > H$ and a prespecified bandwidth H that is large but much smaller than p.

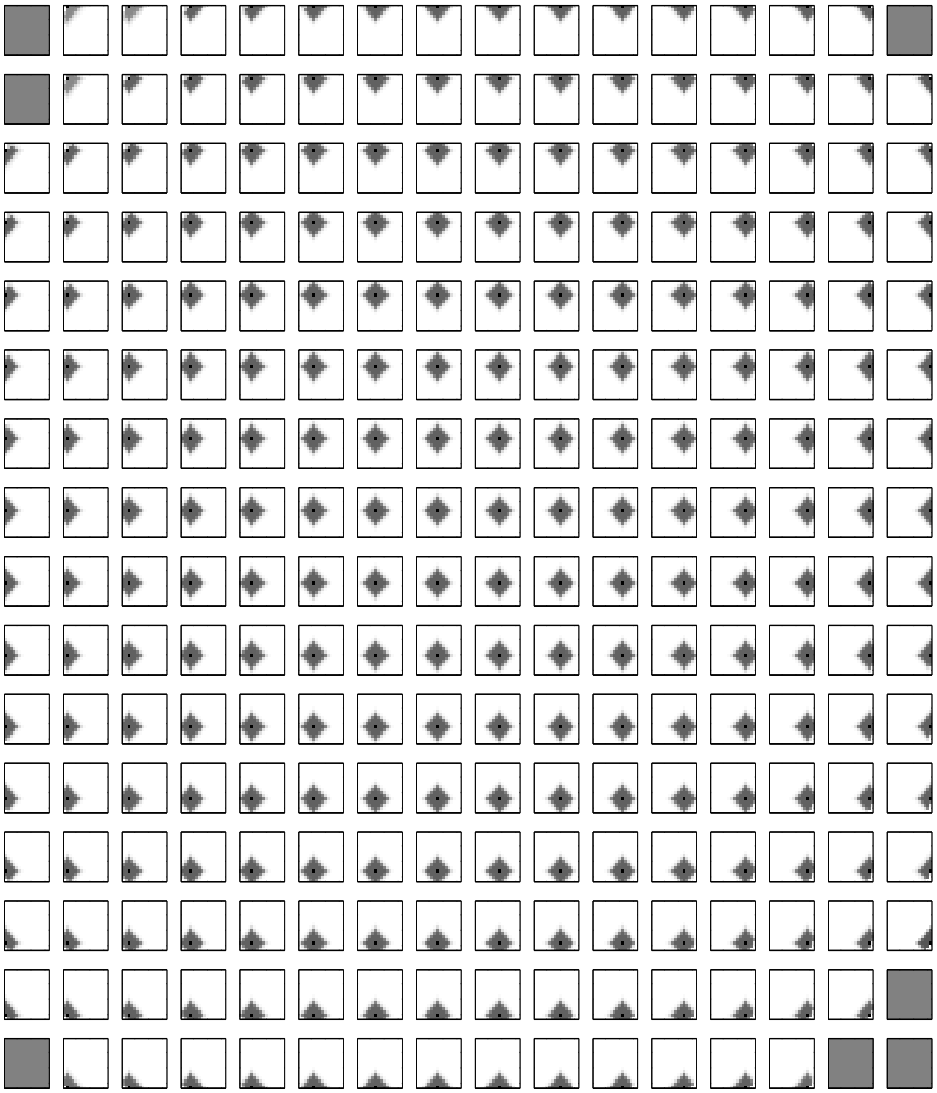


Figure 6: Heatmap plots of percentage of the nonzeros at each location in the estimated inverse covariance matrix from handwritten digit data. Black represents 100%, white 0%.

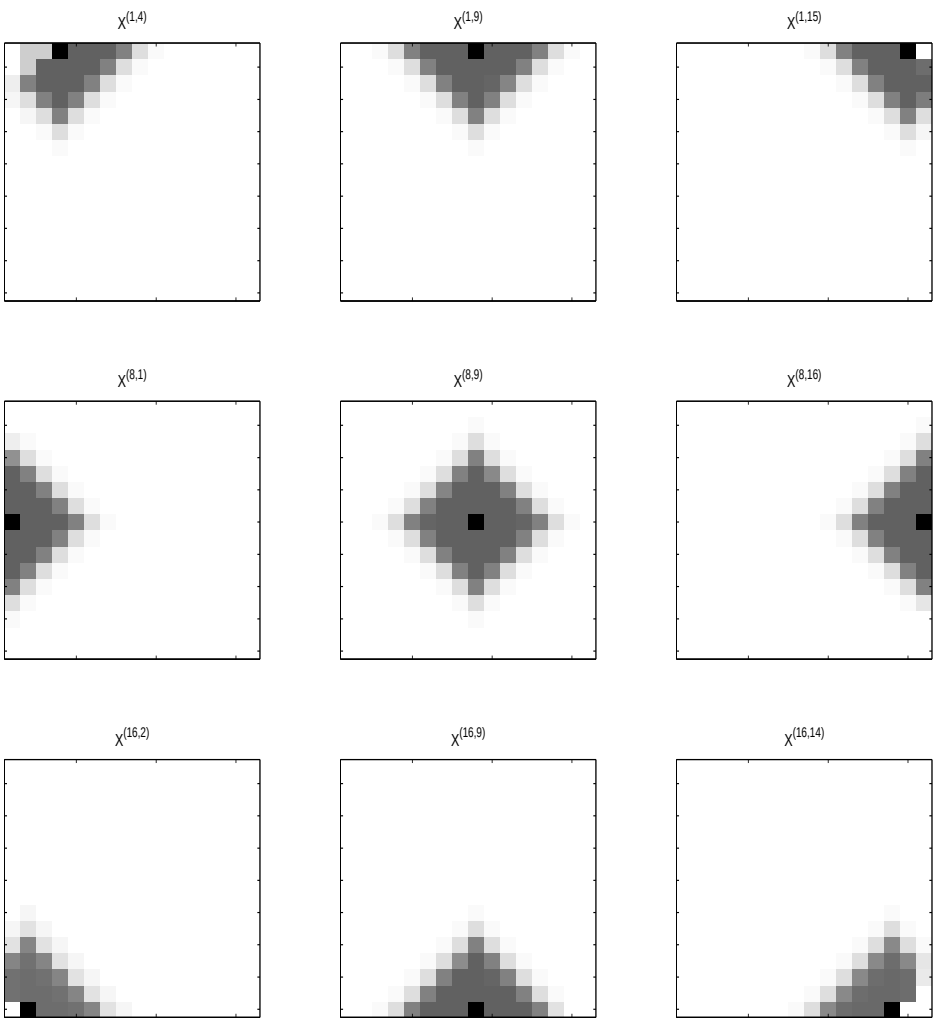


Figure 7: Some of heatmap plots of percentage of the nonzeros at each location in the estimated inverse covariance matrix from handwritten digit data. Black represents 100%, white 0%.

CHAPTER II

A NOTE ON ROBUST KERNEL PRINCIPAL COMPONENT ANALYSIS

2.1 Introduction

Principal component analysis (PCA) is a linear transformation that seeks a coordinate system for a set of multivariate observations such that the greatest variance by any projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on. The new coordinates are referred to as the principal components. By keeping only the first few principal components, PCA achieves dimension reduction while retaining characteristics of the dataset that contribute most to its variation (Jolliffe, 1986).

PCA extracts linear features of high dimensional data. In many applications, however, this can be restrictive and it may be more appropriate to consider nonlinear structures of the data. In recent years, several nonlinear extensions of PCA have been proposed in the literature (Oja, 1982; Hastie and Stuetzle, 1989; Oja, 1991; Bregler and Omohundro, 1994; Schölkopf et al., 1998). In particular, Schölkopf et al. (1998) introduced the kernel PCA. To allow nonlinear features, the kernel PCA performs the classical PCA in a feature space that are nonlinear transformations of the original input variables. Clearly this notion only has conceptual value because the feature space can be of infinite dimension to allow flexible nonlinear features. Nevertheless, Schölkopf et al. (1998) showed that the computation of the kernel PCA only involves the inner product in the feature space. Since the inner product in the feature space can be evaluated through a kernel operator, the kernel PCA can be computed efficiently thanks to the so-called "kernel trick" (Schölkopf and Smola, 2002). The kernel PCA

has seen the explosion of its popularity since its introduction and has proven to be highly successful in various applications such as image analysis, gene expression data analysis among many others.

It is widely recognized that PCA and the kernel PCA can be extremely sensitive to outlying observations, and conclusions drawn based on contaminated principal components can be misleading. Several ways of robustifying the classical PCA have been proposed in the literature (Jackson, 1991). Among many others, these approaches include employing robust estimate of the covariance matrix (Croux and Haesbroeck, 2000) or measure of variation that is more robust than the variance (Ibazizen and Dauxois, 2003). Despite their success in the case of PCA, it is not immediately clear how these approaches can be extended to the kernel PCA.

To fill in this void, we propose a robust kernel PCA in this chapter. Similar to the case of PCA, we use the mean absolute deviation (MAD) to measure the variation by a projection of the data, which is known to be more robust than the variance. We consider applying this robust PCA in the feature space. At the first glance, such a procedure can not be "kernelized" since operations other than inner product are involved in computing MAD. To overcome this problem, we re-formulate our robust kernel PCA using only the inner product in the feature space thanks to the duality property of matrix norms. We also introduce a natural measure to examine the robustness of the original kernel PCA and the proposed robust kernel PCA. We show that this robustness measure can be evaluated using the kernel operator and therefore readily computable for both methods. We use this new measure of influence to show that the robust kernel PCA is much less sensitive to the outlying observations than the original kernel PCA.

The rest of chapter is organized as follows. The methodology of the robust kernel PCA is introduced in the next section. In Section 2.3, we compare the original kernel PCA and the robust kernel PCA based on a perturbation analysis and show that

an outlying observation may have arbitrarily large influence on the original kernel PCA whereas its influence on the robust kernel PCA is always bounded by a constant smaller than one. Section 2.4 presents a simulation study to demonstrate the competitive performance of the robust kernel PCA. To further illustrate the method, we analyze a real data in financial service area using the proposed method in Section 2.5. We conclude with some discussions in Section 2.6.

2.2 Robust Kernel PCA

Given a set of centered observations $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})'$, $k = 1, \ldots, n$, PCA seeks directions that maximize the variance of the projection of the data. For example, the first principal component is given by

$$\arg\max_{\beta} \sum_{k=1}^{n} (\mathbf{x}_k'\beta)^2 \tag{1}$$

It is well known that the variance is extremely sensitive to outliers. To robustify PCA, one can use a more robust measure of variation. In this chapter, we consider using MAD and define our first principal component as

$$\arg\max_{\|\beta\|_2=1} \sum_{k=1}^{n} |\mathbf{x}_k'\beta| \tag{2}$$

To consider nonlinear features of $\mathbf{x}$ that come from a functional space $\mathcal{F}$, we can apply this robust procedure to the basis functions of $\mathcal{F}$, $\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \ldots$. Without loss of generality, assume that $\sum_k \psi_i(\mathbf{x}_k) = 0$ for any i. We look for a vector β of the same dimension as the basis functions such that

$$\beta = \arg\max_{\|\beta\|_2=1} \sum_{k=1}^{n} |\Psi(\mathbf{x}_k)'\beta| \tag{3}$$

where $\Psi(\mathbf{x}) = (\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \ldots)'$.

The functional space $\mathcal{F}$ is often taken to be a reproducing kernel Hilbert space (Wahba, 1990). In such situations, (3) may not be computable since $\mathcal{F}$ can have infinite dimension in a genuine nonparametric setup. A powerful technique to get

around this problem is by the so-called "kernel trick" (Schölkopf and Smola, 2002). Although there are infinitely many basis functions, the inner product in the feature space can always be computed through a kernel operator. The key step therefore is to express the objective in a formulation using only inner products, which is clearly not the case for (3).

To accomplish this goal, we note the duality between the matrix ℓ_p norm and ℓ_q norm given that $1/p + 1/q = 1$. Simple derivation leads to the following matrix transposition invariant property (Choulakian, 2005). Let A be a $m \times n$ matrix, define

$$\|A\|_{pr} = \max_{\|x\|_r=1:\mathbf{x}\in R^n} \|A\mathbf{x}\|_p, \tag{4}$$

where $\|\cdot\|_p$ is a vector p-norm and $p,\ r > 0$. The transposition invariant property states that

$$\|A\|_{pr} = \|A'\|_{sq}, \tag{5}$$

where

$$\frac{1}{p} + \frac{1}{q} = 1, \qquad \frac{1}{r} + \frac{1}{s} = 1. \tag{6}$$

Now define $A_{ij} = \psi_j(\mathbf{x}_i)$. Then an application of (5) implies that

$$\max_{\|\beta\|_2=1} \sum_{k=1}^{n} |\Psi(\mathbf{x}_k)'\beta| = \|A\|_{12} = \|A'\|_{2\infty} = \max_{\|\alpha\|_\infty=1} \sqrt{\alpha' AA'\alpha} \tag{7}$$

Note that the (i, j) entry of AA' is $\langle \Psi(\mathbf{x}_i), \Psi(\mathbf{x}_j)\rangle = K(\mathbf{x}_i, \mathbf{x}_j)$. To evaluate the right hand side of (7), it is sufficient to know the kernel operator $K(\cdot, \cdot)$. This kernel representation allows us to compute the value of the inner product in $\mathcal{F}$ without having to carry out the map Ψ. This method was previously used by Boser et al. (1992) to extend the Generalized Portrait hyperplane classifier of Vapnik and Chervonenkis (1974) to nonlinear support vector machines by substituting a pre-specified kernel function $K(\cdot, \cdot)$ for all occurrences of inner products. The readers are referred to Schölkopf and Smola (2002) for a detailed account of this so-called "kernel trick". It is also known that there is a one-to-one correspondence between a reproducing kernel

Hilbert space and a positive definite kernel operator $K(\cdot,\cdot)$. For this reason, it is often times convenient to directly specify the kernel operator instead of the functional space itself. Kernels that are commonly used in practice include the polynomial kernels and Gaussian kernels.

The polynomial kernel of degree d is given by

$$K(\mathbf{x},\mathbf{y}) = (\langle \mathbf{x},\mathbf{y}\rangle + 1)^d . \tag{8}$$

Besides the polynomial kernel, Gaussian kernel

$$K(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{\sigma^2}\right) \tag{9}$$

is also very popular.

With slight abuse of notation, denote K a $n \times n$ matrix whose (i,j) entry is $K(\mathbf{x}_i,\mathbf{x}_j)$. Then we can rewrite the right hand side of (7) as

$$\widehat{\alpha}^{(1)} = \arg \max_{\|\alpha\|_\infty = 1 : \alpha \in R^n} \sqrt{\alpha' K \alpha} = \arg \max_{\|\alpha\|_\infty = 1 : \alpha \in R^n} \alpha' K \alpha, \tag{10}$$

where the superscript is used to indicate that it corresponds to the first principal component.

Once $\alpha^{(1)}$ is obtained, again by the transposition invariant property, the maximizer of the left hand side of (7) is given by $\widehat{\beta}^{(1)} = A'\alpha^{(1)}/\sqrt{(\alpha^{(1)})' K \alpha^{(1)}}$, which again requires the knowledge of map Ψ. Fortunately, we are only interested in the projection of a data point $\mathbf{x}$ into the principal components, which can be computed as

$$\Psi(\mathbf{x})'\beta^{(1)} = \frac{\Psi(\mathbf{x})' A' \alpha^{(1)}}{\sqrt{(\alpha^{(1)})' K \alpha^{(1)}}} = \frac{\sum_{k=1}^n \alpha_k^{(1)} \langle \Psi(\mathbf{x}), \Psi(\mathbf{x}_k)\rangle}{\sqrt{(\alpha^{(1)})' K \alpha^{(1)}}} = \frac{\sum_{k=1}^n \alpha_k^{(1)} K(\mathbf{x},\mathbf{x}_k)}{\sqrt{(\alpha^{(1)})' K \alpha^{(1)}}}. \tag{11}$$

After the first principal component is obtained, we then target at the second principal component which is orthogonal to the first one. We first project the data from the feature space $\mathcal{F}$ into its linear subspace that is orthogonal to the first principal component. Note that the second principal component is now the first principal component of the projected data. The aforementioned procedure for the first principal

component can then be applied if we know how to compute the kernel operator in the linear subspace. Let $\Psi(\mathbf{x})$ be a point in $\mathcal{F}$, then $\Psi(\mathbf{x}) - \widehat{\beta}^{(1)}(\widehat{\beta}^{(1)})'\Psi(\mathbf{x})$ is its projection into the linear subspace that is orthogonal to $\widehat{\beta}^{(1)}$. The inner product of the linear subspace can be calculated:

$$\begin{aligned} K^{(2)}(\mathbf{x},\mathbf{y}) &= \langle \Psi(\mathbf{x}) - \widehat{\beta}^{(1)}(\widehat{\beta}^{(1)})'\Psi(\mathbf{x}), \Psi(\mathbf{y}) - \widehat{\beta}^{(1)}(\widehat{\beta}^{(1)})'\Psi(\mathbf{y})\rangle \\ &= \langle \Psi(\mathbf{x}), \Psi(\mathbf{y})\rangle - 2\langle \Psi(\mathbf{x}), \widehat{\beta}^{(1)}(\widehat{\beta}^{(1)})'\Psi(\mathbf{y})\rangle \\ &\quad + \langle \widehat{\beta}^{(1)}(\widehat{\beta}^{(1)})'\Psi(\mathbf{x}), \widehat{\beta}^{(1)}(\widehat{\beta}^{(1)})'\Psi(\mathbf{y})\rangle \\ &= \langle \Psi(\mathbf{x}), \Psi(\mathbf{y})\rangle - \langle \Psi(\mathbf{x}), \widehat{\beta}^{(1)}(\widehat{\beta}^{(1)})'\Psi(\mathbf{y})\rangle \end{aligned} \tag{12}$$

Note that $\widehat{\beta}^{(1)} = A'\alpha^{(1)}/\sqrt{(\alpha^{(1)})'K\alpha^{(1)}}$

$$\langle \Psi(\mathbf{x}), \widehat{\beta}^{(1)}(\widehat{\beta}^{(1)})'\Psi(\mathbf{y})\rangle = \frac{\Psi(\mathbf{x})'A'\alpha^{(1)}(\alpha^{(1)})'A\Psi(\mathbf{y})}{(\alpha^{(1)})'K\alpha^{(1)}} = \frac{\sum_{i,j=1}^{n}\alpha_i^{(1)}\alpha_j^{(1)}K(\mathbf{x},\mathbf{x}_i)K(\mathbf{x}_j,\mathbf{y})}{(\alpha^{(1)})'K\alpha^{(1)}}. \tag{13}$$

Therefore,

$$K^{(2)}(\mathbf{x},\mathbf{y}) = K(\mathbf{x},\mathbf{y}) - \frac{\sum_{i,j=1}^{n}\alpha_i^{(1)}\alpha_j^{(1)}K(\mathbf{x},\mathbf{x}_i)K(\mathbf{x}_j,\mathbf{y})}{(\alpha^{(1)})'K\alpha^{(1)}} \tag{14}$$

which can be computed without knowing Ψ.

The rest of the principle components can be computed in a similar fashion. In general, the kernel operator needed for the rth principal component is

$$K^{(r)}(\mathbf{x},\mathbf{y}) = K(\mathbf{x},\mathbf{y}) - \langle \Psi(\mathbf{x}), W\Psi(\mathbf{y})\rangle \tag{15}$$

where $W = (\beta^{(1)},\ldots,\beta^{(r-1)})(\beta^{(1)},\ldots,\beta^{(r-1)})'$.

To sum up, our proposed robust kernel PCA method can be computed using the following recipe:

2.3 Perturbation Analysis

The influence function is a commonly used measure of the robustness for a statistical procedure. The influence function of a statistical functional $T_0(F)$ is defined as

$$IC_{T_0,F}(z) = \lim_{\epsilon\to 0}\frac{T_0(F_\epsilon) - T_0(F)}{\epsilon} = \left.\frac{\partial T_0(F_\epsilon)}{\partial\epsilon}\right|_{\epsilon=0} \tag{16}$$

Algorithm 1 Compute First R Robust Kernel Principal Components

Step 1. Compute $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ for all $i, j = 1, \ldots, n$.

Step 2. Center the kernel matrix $\bar{K} = K - \mathbf{1}K/n - K\mathbf{1}/n + \mathbf{1}K\mathbf{1}/n^2$, where $\mathbf{1}$ is a $n \times n$ matrix with ones.

Step 3. Compute the first principal component through α using (10) and kernel $\bar{K}$.

Step 4. For $r = 2$ to R

(a) Compute the kernel matrix $(K^{(r)}(\mathbf{x}_i, \mathbf{x}_j))$ using (15)

(b) Center the kernel matrix as in Step 2.

(c) Compute the rth principal component using (10) with the kernel matrix obtained

---

where F is a distribution function, $F_\epsilon = (1-\epsilon)F + \epsilon\delta_z$ and δ_z is a point mass at z. Of particular interest is the choice of $z = \mathbf{x}_i$, $\epsilon = 1/(n-1)$ and F being the empirical distribution function, which amounts to measuring the influence of deleting the ith case (Cook and Weisberg, 1982). Instead of deleting cases one at a time, some authors have suggested to perturb a single case and the influence of the corresponding case is investigated through the deriviative of the perturbation. To formalize this approach, assign each case a weight w_i $(i = 1, \ldots, n)$. Denote $T_{\mathbf{w}}$ the statistic with weights $\mathbf{w} = (w_1, \ldots, w_n)'$. The influence of the ith case is given as

$$\left.\frac{\partial T_{\mathbf{w}}}{\partial w_i}\right|_{\mathbf{w}=(1,\ldots,1)'} \tag{17}$$

In the case of the kernel PCA, let β be a principal component in $\mathcal{F}$. The influence of the ith observation on the projection of a future data point $\mathbf{x}_0$ is

$$\Psi(\mathbf{x}_0)' \left.\frac{\partial \beta}{\partial w_i}\right|_{\mathbf{w}=(1,\ldots,1)'}. \tag{18}$$

It is therefore natural to measure the robustness of β using

$$\mathrm{IF}_i(\beta) = \left|\left| \left.\frac{\partial \beta}{\partial w_i}\right|_{\mathbf{w}=(1,\ldots,1)'} \right|\right|_2^2 \tag{19}$$

To fix ideas, we consider only the first principle component $\widehat{\beta}^{(1)}$ in the following discussion. We begin with the original kernel PCA of Schölkopf et al. (1998). Note

that $\widehat{\beta}^{(1)}$ is the linear principal component in $\mathcal{F}$, Critchley (1985) has shown that

$$\mathrm{IF}_i(\widehat{\beta}^{(1)}) = \left(\frac{2}{n}\right)^2 \left((\Psi(\mathbf{x}_i)'\widehat{\beta}^{(1)}\right)^2 \sum_{r>1} \frac{\left(\Psi(\mathbf{x}_i)'\widehat{\beta}^{(r)}\right)^2}{(\widehat{\lambda}_1 - \widehat{\lambda}_r)^2}, \tag{20}$$

where $\widehat{\lambda}_r'$s are the eigenvalues corresponding to $\widehat{\alpha}^{(r)}$. Clearly, the influence function is unbounded for certain outlying observations.

To evaluate this influence function, we need to compute $\Psi(\mathbf{x}_i)'\widehat{\beta}^{(r)}$, the projection of the ith observation on the rth principal component. To this end, we apply the transposition invariant property to (1),

$$\max_{\beta} \sum_{k=1}^{n} (\mathbf{x}_k'\beta)^2 = \max_{\alpha} \alpha' K \alpha. \tag{21}$$

Therefore, $\widehat{\beta}^{(1)} = A'\widehat{\alpha}^{(1)}/\sqrt{(\widehat{\alpha}^{(1)})'K\widehat{\alpha}^{(1)}}$ where $\widehat{\alpha}^{(1)}$ is the first principal component of K. Now the influence function can be re-written in terms of $\widehat{\alpha}^{(1)}$:

$$\mathrm{IF}_i(\widehat{\beta}^{(1)}) = \left(\frac{2}{n}\right)^2 \frac{\left(K_i\widehat{\alpha}^{(1)}\right)^2}{(\widehat{\alpha}^{(1)})'K\widehat{\alpha}^{(1)}} \sum_{r>1} \frac{\left(K_i\widehat{\alpha}^{(r)}\right)^2}{(\widehat{\alpha}^{(r)})'K\widehat{\alpha}^{(r)}(\widehat{\lambda}_1 - \widehat{\lambda}_r)^2} \tag{22}$$

where K_i is the ith row of K. Note that (22) can be computed without knowing the map Ψ.

In the case of robust kernel PCA, after introducing the weights $w_1, \ldots, w_n$, we can rewrite (10) as

$$\widehat{\alpha}^{(1)*} = \arg \max_{\|\alpha\|_\infty = 1 : \alpha \in R^n} \alpha' \Omega K \Omega \alpha, \tag{23}$$

where Ω is a diagonal matrix whose (i,i) entry is w_i. Because of the discrete nature of the feasible set, $\widehat{\alpha}^{(1)*} = \widehat{\alpha}^{(1)}$ given that w_i's are sufficiently close to 1. Therefore, after perturbation,

$$\widehat{\beta}^{(1)*} = \frac{A'\Omega\alpha^{(1)*}}{\sqrt{(\alpha^{(1)*})'\Omega K \Omega \alpha^{(1)*}}} = \frac{A'\Omega\alpha^{(1)}}{\sqrt{(\alpha^{(1)})'\Omega K \Omega \alpha^{(1)}}} \tag{24}$$

Let $w_i = 1 - \epsilon$ and $w_j = 1$ for all $j \neq i$

$$\widehat{\beta}^{(1)*} = \widehat{\beta}^{(1)} - \frac{\epsilon}{\sqrt{(\alpha^{(1)})'K\alpha^{(1)}}} \left(A'H_i\alpha^{(1)} - \frac{(\alpha^{(1)})'KH_i\alpha^{(1)}}{(\alpha^{(1)})'K\alpha^{(1)}} A'\alpha^{(1)}\right) + O(\epsilon^2), \tag{25}$$

where H_i is a $n \times n$ matrix with zeros except that its (i,i)th entry is one. It is natural to define the influence of perturbing the ith observation as

$$\begin{aligned} \mathrm{IF}_i(\widehat{\beta}^{(1)}) &= \left|\left| \frac{1}{\sqrt{(\alpha^{(1)})' K \alpha^{(1)}}} \left(A' H_i \alpha^{(1)} - \frac{(\alpha^{(1)})' K H_i \alpha^{(1)}}{(\alpha^{(1)})' K \alpha^{(1)}} A' \alpha^{(1)} \right) \right|\right|_2^2 \qquad (26) \\ &= \frac{(\alpha^{(1)})' H_i K H_i \alpha^{(1)}}{(\alpha^{(1)})' K \alpha^{(1)}} - \frac{\left((\alpha^{(1)})' K H_i \alpha^{(1)}\right)^2}{\left((\alpha^{(1)})' K \alpha^{(1)}\right)^2} \qquad (27) \end{aligned}$$

In contrast to the original kernel PCA, the influence function of the robust kernel PCA is bounded by the first term. To be specific, note that $\|H_i \alpha^{(1)}\|_\infty = \alpha_i^{(1)} \leq \|\alpha^{(1)}\|_\infty = 1$. We have

$$\mathrm{IF}_i(\widehat{\beta}^{(1)}) \leq \frac{(\alpha^{(1)})' H_i K H_i \alpha^{(1)}}{(\alpha^{(1)})' K \alpha^{(1)}} < 1, \qquad (28)$$

from the definition of $\alpha^{(1)}$.

2.4 Simulation

To illustrate the methodology, we first consider a toy example. We use this example to demonstrate the robustness of the proposed approach. We first randomly generate fifty data points around a circle in the two dimensional space. Each point is generated in the following fashion. First an angle is sampled from a uniform distribution between 0 and 2π. The radius is then randomly generated from $N(3, 0.05^2)$. In the top panels of Figure 8 we plot the data points together with the first original kernel principal component and first robust kernel principal component. We use the polynomial kernel with degree two for both methods. The two methods perform very similarly in this case. Now we add an outlier to the data. The outlying observation is located at $(5,5)$. We plot in the bottom panels of Figure 8 the first original kernel principal component and the first robust kernel principal component of the contaminated data. The result suggests that the influence of the outlier on the original kernel principal component is quite significant, but marginal for the robust kernel principal component.

Next, we examine the robustness property from a different angle by looking at the perturbation analysis from Section 3. For each method, we compute the influence of

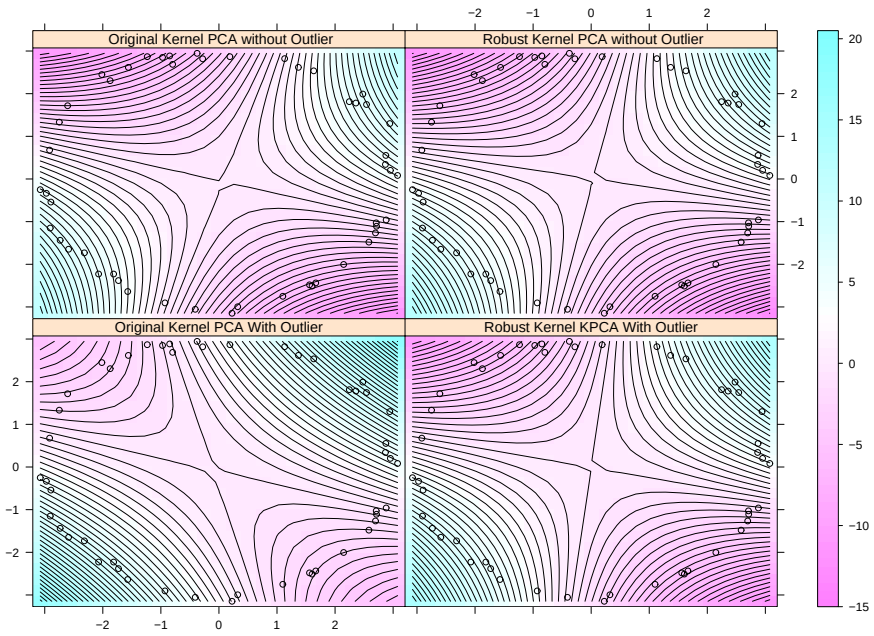


Figure 8: First kernel principal component for the two-dimensional circle example

each observation. To make the influence measure comparable in magnitude for the two different methods, we scale the influence values so that the sum of the influence over all observations is one. We compare the normalized influence of the outlier for the two methods. The comparison is based on 1000 datasets simulated in the aforementioned fashion. The pairwise comparison of the influence is given in Figure 9, from which we see a significant reduction of the influence of the outlier for our robust kernel PCA.

2.5 Real Example

We now apply our method to a real application in financial service. For the purpose of surveillance, it is of great importance to characterize the normal transaction behavior in contrast with the suspicious ones. The banking experts often times look over several important aspects of an account history such as the number of transactions, the total amount of transactions among others in order to reveal transaction patterns. In a particular example, the experts suspect that there might be suspicious cases among a

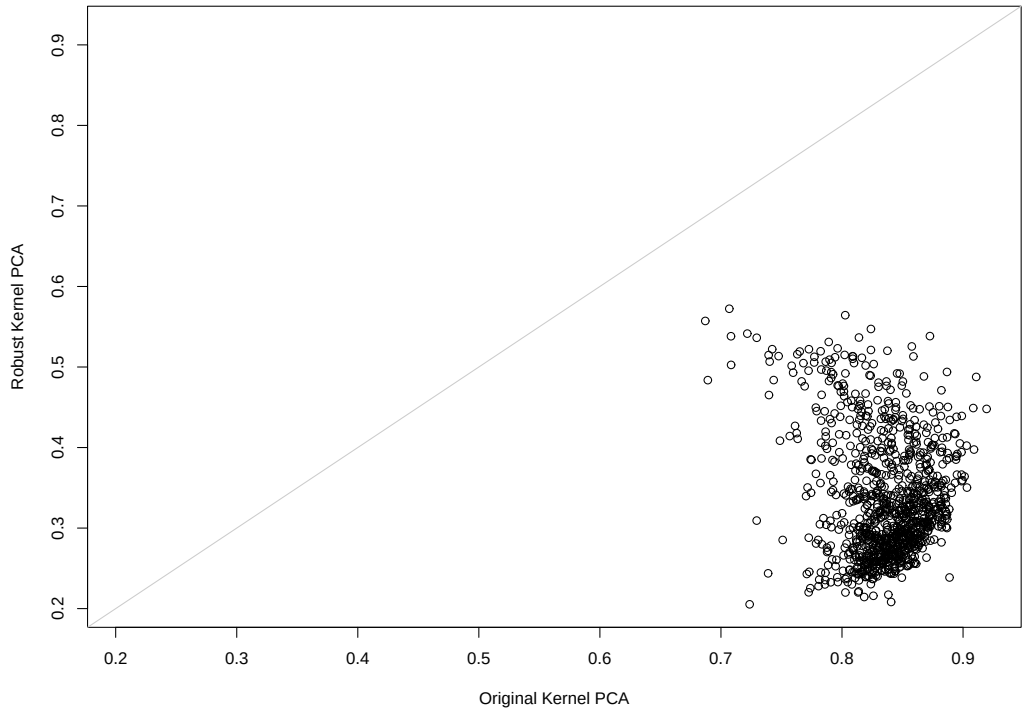
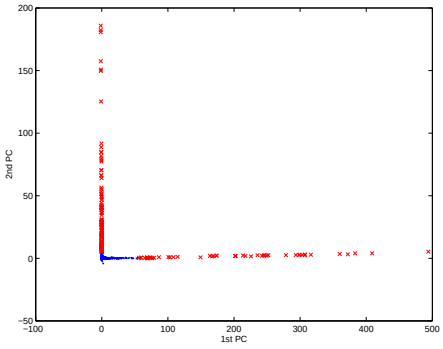


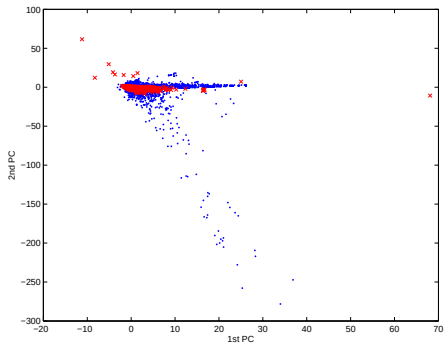
Figure 9: Influence measure of the outlier for the two-dimensional circle example

sample of 6321 accounts. The account history in a one-month period is summarized by eight statistical measurements. For confidentiality reason, we do not disclose more details about the data we are using here. It is clearly very time-consuming for the expert to look over all the cases. Efficient dimension reduction and visualization tool such as the kernel PCA would prove extremely helpful in this aspect. We apply the original kernel PCA and robust kernel PCA on the data to extract the first two kernel principal components and project all cases in a two dimensional space spanned by these components. We immediately see from both plots that there might be outlying cases, or in other words abnormal behavior in this dataset. The question next is which method more accurately characterizes the normal pattern and identifies suspicious activities. To this end, we re-run both kernel PCA methods with the corresponding outlying observations removed and project all cases in the new principal component space. The projections are given in the following figure with the red dots representing

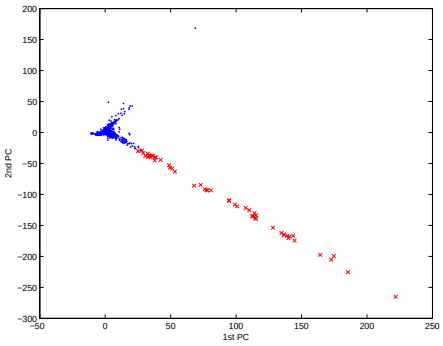
the outliers identified by each method. For the original kernel PCA, majority of the outlying observations found in the original analysis do not appear to be abnormal anymore. One plausible explanation is that the original kernel principal components found on the original data were influenced by the truly abnormal cases and the two-dimensional projection fails to capture the real pattern of the normal activities. In contrast the outlying observations found by robust kernel PCA still appear to be abnormal.



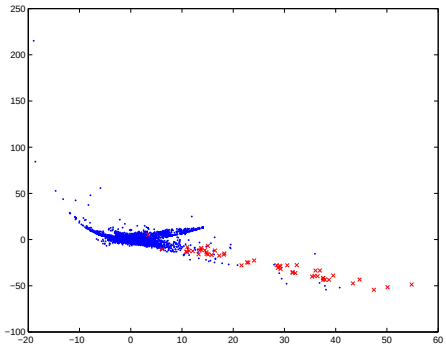
(a) Original Kernel PCA with "Outlier"



(b) Original Kernel PCA without "Outlier"



(c) Robust Kernel PCA with "Outlier"



(d) Robust Kernel PCA without "Outlier"

2.6 Conclusion

It is known that the kernel PCA may suffer from the presence of outlying observations. Taking advantage of the dual matrix norms, we propose a robust kernel PCA procedure in this chapter. We demonstrate by a simulation study and a real application that the proposed method is more robust to outliers than the original kernel

PCA.

A more general class of principal direction can be given in the feature space as

$$\arg \max_{\|\beta\|_2=1} \|A'\beta\|_p, \tag{29}$$

for some $1 \leq p \leq 2$. The original kernel PCA takes $p = 2$ whereas our robust kernel PCA chooses $p = 1$. Although we have focused on using the mean absolute deviation in this note, it is worth noting that all these kernel PCA can be "kernelized" in the same fashion as our robust kernel PCA. In particular, they are also determined by a n dimensional vector

$$\arg \max_{\|\alpha\|_q=1} \alpha' K \alpha, \tag{30}$$

where q is such that $1/p + 1/q = 1$. We choose $p = 1$ because of the robustness it brings about. Other choices may also have their own merits. We leave this for future studies.

CHAPTER III

ACTIVE LEARNING VIA SEQUENTIAL DESIGN WITH APPLICATIONS TO DETECTION OF MONEY LAUNDERING

3.1 Introduction

Money laundering is an act to hide the true origin of funds by sending them through a series of seemingly legitimate transactions. Its main purpose is to conceal the fact that funds were acquired as a result of some form of criminal activity. These laundered funds can in turn be used to foster further illegal activities such as the financing of terrorist activity or trafficking of illegal drugs. Even legitimate funds that are laundered to avoid reporting them to the government, as is the case with tax evasion, lead to substantial costs for society. Financial institutions which have the responsibility to detect and prevent money laundering are facing a challenge to sort through potential suspicious activities among millions of legitimate transactions every day. Once suspicious activities have been detected, an investigation effort can easily take 10 hours to classify a case as suspicious or non-suspicious. Figure 10 shows a sample of transaction data. There are all kinds of information in the transaction history. Therefore, investigating every account to detect money laundering is extremely time-consuming and cost prohibitive.

One way to overcome this problem is to extract certain statistical features based on the transaction history of each account. If these statistical features are highly representative for the suspiciousness for the transaction history, then they can be used to prioritize the accounts for investigation. The accounts with high priority are investigated thoroughly to find their suspiciousness level.

Acct No.	D/C	PostDate	TransAmt	TransCode	Description
999999	D	1/23/2005	$1,295.00	9059	Check Check
999999	D	5/19/2004	$1,020.00	9059	Check Check
999999	D	1/23/2005	$10,000.00	9059	Check Check
999999	D	3/2/2004	$5.00	9593	Returned Item Charge Returned Item Charge
999999	D	2/24/2004	$5.00	9593	Returned Item Charge Returned Item Charge
999999	D	10/12/2004	$34.00	9203	Overdraft Charge Overdraft Charge
999999	D	7/13/2004	$60.00	9659	Check Card Purchase Dr Jm Layton And Ep Lay5194121949512823
999999	D	6/10/2004	$129.36	9905	Pos Withdrawal Costco Whse #0001 8442627516108999910830
999999	D	6/14/2004	$51.49	9905	Pos Withdrawal Bed, Bath & Beyo 84426275165089999914310
999999	D	6/10/2004	$168.44	9905	Pos Withdrawal Costco Whse #0001 8442627516108999910370
999999	D	7/18/2004	$34.84	9905	Pos Withdrawal Costco Whse #0001 8442627519708999916890
999999	D	5/24/2004	$33.20	9905	Pos Withdrawal Costco Gas #00662 8442627514408999924800
999999	D	6/22/2004	$158.65	9905	Pos Withdrawal Bed, Bath & Beyo 8442627517308999922610
999999	D	6/10/2004	$190.64	9905	Pos Withdrawal Costco Whse #0001 8442627516108999910750
999999	C	1/14/2004	$100.00	9003	Deposit Deposit
999999	C	8/10/2004	$20.00	9003	Deposit Deposit
999999	C	5/11/2004	$10,000.00	9003	Deposit Deposit
999999	C	8/31/2004	$3,300.00	9003	Deposit Deposit 0831CA319P007160134679
999999	C	6/29/2004	$2,079.95	9003	Deposit Deposit
999999	C	10/7/2004	$2,500.00	9003	Deposit Deposit
999999	C	1/30/2005	$22.43	9699	Automatic Deposit Deposit Merchant Bankcd 267917678885
999999	C	1/30/2005	$22.43	9699	Automatic Deposit Deposit Merchant Bankcd 267917678885
999999	C	6/16/2004	$64.97	9660	Reverse Check Card Purchase The Home Depot 4715 5166010183470016
999999	C	7/21/2004	$151.61	9660	Reverse Check Card Purchase Hardware Sales 5202207788501885
999999	C	9/20/2004	$24.95	9660	Reverse Check Card Purchase Twx*Sports Illustrated 5259000879500624
999999	C	4/27/2004	$14,032.37	9039	Deposit To Close Account Deposit To Close Account
999999	C	11/30/2004	$3,243.59	9003	Deposit Deposit
999999	C	7/6/2004	$400.00	9003	Deposit Deposit
999999	C	10/6/2004	$2,981.07	9003	Deposit Deposit
999999	C	7/21/2004	$100.00	9007	Miscellaneous Deposit Transfer From Checking 22782403

Figure 10: A sample of transaction data

The problem can be formulated as follows. Let $\boldsymbol{x} = (x_1, \ldots, x_p)'$ be the vector of feature variables extracted from a transaction history. Let $Y = 1$ if the account is classified as suspicious and $Y = 0$ otherwise. Then, $P(Y = 1|\boldsymbol{x}) = F(\boldsymbol{x})$ gives the probability of suspiciousness at a given level of $\boldsymbol{x}$. When $F(\boldsymbol{x})$ exceeds a threshold probability α, we can investigate that account in detail. Assume that $F(\boldsymbol{x})$ is an increasing function in each x_i. Define the threshold hyperplane $l_{\boldsymbol{x}}$ at level α as

$$l_{\boldsymbol{x}} = \{\boldsymbol{x} : F(\boldsymbol{x}) = \alpha\}. \tag{1}$$

Now for a new account, if $\boldsymbol{x}$ falls below $l_{\boldsymbol{x}}$, then we need not investigate that account further. But if $\boldsymbol{x}$ falls above $l_{\boldsymbol{x}}$, we must investigate the account in detail. An institution may choose a reasonable α so that only a portion of accounts needs to be investigated. This scientific approach can significantly improve productivity by investigating cases that really matter.

The challenge is not only due to the huge amount of transactions each day, but also due to different kinds of business with money laundering activities. The behaviors of various business categories can be quite different. Even the behaviors of the same business category at different time periods appear to be different in money laundering

activities. For example, given a specified suspicious level α, the threshold hyperplane for personal accounts can be completely different from that of small business accounts. Even the importance of statistical features can vary dramatically. Thus, when a new set of accounts is introduced, it is not likely to share the same threshold hyperplane from the past investigation.

It is important to develop a procedure for finding the threshold hyperplane efficiently. The problem is that $F(\boldsymbol{x})$ is unknown and therefore, $l_{\boldsymbol{x}}$ is also unknown. Data on $\boldsymbol{x}$ and Y can be used to estimate $l_{\boldsymbol{x}}$. For this purpose, a training set of the investigated accounts is needed. However, labelling the suspiciousness (1 or 0) for a large number of accounts is time consuming and extremely expensive. It will be beneficial to find a way to minimize the number of investigated accounts and use them to construct effective threshold hyperplane. This calls for active learning methods (Mackay, 1992; Cohn et al., 1996; Fukumizu, 2000). Here, the learner actively selects data points to be added into the training set. *In this chapter, an active learning method that improves the process of money laundering detection is proposed.*

The remaining part of the chapter is organized as follows. In section 3.2, we give the motivation for the proposed active learning method using sequential designs. Section 3.3 reviews some existing methods in sequential designs and the concept of optimal designs. The active learning via sequential design is proposed in Section 3.4. In Section 3.5, we implemented the proposed method into a real case study for detecting money laundering. Section 3.6 presents some simulation results to demonstrate the performance of the proposed active learning approach. Some discussions and conclusions are given in Section 3.7.

3.2 Motivation

To minimize the number of investigated accounts and use them to construct effective threshold hyperplane, we need to judiciously select the accounts for investigation.

This call for the use of active learning in machine learning. Recently, active learning methods using support vector machines (SVM) were developed by several researchers (Tong and Koller, 2001; Schohn and Cohn, 2000; Campbell et al., 2000), which can be applied to the present problem.

For binary response, active learning with SVM is mainly for two-class classification. The decision boundary in SVM implements the Bayes rule $P(Y|\boldsymbol{x}) = 0.5$, which is a special case of (1). In money laundering detection, sometimes the interest lies in values other than $\alpha = 0.5$. It is important to find the threshold hyperplane at a higher value of α such as $\alpha = 0.75$. To address this point, we propose a new active learning method using sequential designs. The sequential nature of the method helps to identify the suspicious accounts with reasonable time and effort. In statistical design of experiments, the locations of training data points are chosen by the users so as to maximize the information in the experiment (Kiefer, 1959; Fedorov, 1972; Pukelsheim, 1993). In sequential designs, the training data points are selected sequentially, i.e., the next point to be selected for training is based on information gathered for previously trained data points. However, in money laundering detection, the problem is different from classic sequential design. Because the accounts are already available, we cannot select arbitrary setting of accounts to get investigation response. Noting that the motivation of active learning in machine learning is closely related to sequential designs in statistics, in this chapter, we will exploit the synergies between these two approaches to develop a new active learning procedure based on sequential designs and optimal designs. It provides a more flexible way to get threshold hyperplane for different values of α.

3.3 Review of Sequential Designs

The problem of estimating the threshold hyperplane is closely related to the problem of stochastic root-finding. Suppose we want to find the root of an unknown

univariate function $E(Y|x) = F(x)$. The root can be estimated from the data $(x_1, Y_1), \ldots, (x_n, Y_n)$. In sequential designs, the data points are chosen sequentially, i.e., x_{n+1} is selected based on $x_1, x_2, \ldots, x_n$ and their corresponding response $Y_1, Y_2, \ldots, Y_n$. There are two approaches to generating sequential designs: stochastic approximation and optimal design.

In stochastic approximation methods, the x's are chosen such that x_n converges to the root as $n \to \infty$. Robbins and Monro (1951) proposed the stochastic approximation procedure given by

$$x_{n+1} = x_n - a_n(Y_n - \alpha), \tag{2}$$

where $\{a_n\}$ is a pre-specified sequence of positive constants. They also established the conditions under which x_n converges to the root. This stochastic approximation method is also one of the classical pattern classification methods (Duda et al., 2001). An interesting modification of the Robbins-Monro procedure for binary data was proposed by Joseph (2004). Wu (1985) proposed another stochastic approximation method known as the "logit-MLE method", in which $F(x)$ is approximated by a parametric function $H(x|\boldsymbol{\theta})$. Then, determination of x_{n+1} is a two-step procedure. First, a maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ is found from $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$. Then x_{n+1} is chosen as $H(x_{n+1}|\hat{\boldsymbol{\theta}}_n) = \alpha$. Ying and Wu (1997) showed the convergence of x_n almost surely irrespective of the function $F(x)$. Because of the efficient utilization of the complete data, the logit-MLE performs better than the Robbins-Monro procedure. Joseph et al. (2007) proposed a stochastic approximation method that gives more weights to data points closer to the root via a Bayesian scheme.

In the optimal design approach to sequential designs, first a parametric model for the unknown function is postulated. Then, the x points are chosen sequentially based on some optimality criteria (Kiefer, 1959; Fedorov, 1972; Pukelsheim, 1993). For example, Neyer (1994) proposed a sequential D-optimality-based design. Here x_{n+1} is chosen so that the determinant of the estimated Fisher information is maximized.

It is well known that a D-optimal criterion minimizes the volume of the confidence ellipsoid of the parameters (Silvey, 1980). The root is solved from the final estimate of the function $F(x)$.

The performance of the optimal design approach is model dependent. It performs best when the assumed model is the true model, but the performance deteriorates as the model deviates from the true model. One attractive property of the stochastic approximation methods, including the logit-MLE, is the robustness of their performance to model assumptions. This is because as n becomes large, the points get clustered around the root which enable the estimation of root irrespective of the model assumption. Understandably, the performance of the stochastic approximation method is not as good as the optimal design when the assumed model in the latter approach is valid. This point was confirmed by Young and Easterling (1994) through extensive simulations. In this chapter, we propose a new sequential design approach that combines the advantages of both approaches. Our approach is expected to be robust to model assumptions as in stochastic approximation methods as well as produce comparable performance to optimal design approach when the model assumptions are valid.

One shortcoming of the aforementioned methods is that they can only be applied to univariate problems. But in money laundering detection example and other applications (e.g., junk email classification), more than one statistical feature of the data are of interest. Therefore, it is important to extend the existing methods to multivariate problems. We propose a simple approach to account for the multivariate nature of the data. The methodology is explained in the next section.

3.4 Methodology

3.4.1 Active Learning via Sequential Design

In pool-based active learning (Lewis and Gale, 1994), there is a pool of unlabelled data. The learner has access to the pool and can request the true label for a certain

number of data in the pool. The main issue is in finding a way to choose the next unlabelled data point to get the response. The proposed active learning via sequential design attempts to get "close in" on the region of interest efficiently, meanwhile improves the estimation accuracy of $l_{\boldsymbol{x}}$ for a given α.

For the ease of exposition, we explain the methodology with two variables $\boldsymbol{x} = (x_1, x_2)^T$. It can be easily extended to more than two variables. We assume that each variable has a positive relationship with the response, i.e., for larger value of x_j, the probability is higher to get the response $Y = 1$. Define a *synthetic variable* z by $z = wx_1 + (1-w)x_2$, where w is an *unknown* weight factor in $[0, 1]$. By doing this we can convert the multivariate problem into a univariate problem, so that the existing methods for sequential designs can be easily applied.

As in the case of Wu's logit-MLE method, assume the model

$$F(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{e^{(z-\mu)/\sigma}}{1 + e^{(z-\mu)/\sigma}}, \tag{3}$$

which has three parameters $\boldsymbol{\theta} = (\mu, \sigma, w)^T$. As noted before, its convergence is independent of the logit model assumption. By the definition in (1), the threshold hyperplane $l_{\boldsymbol{x}}$ is

$$l_{\boldsymbol{x}} = \{\boldsymbol{x} = (x_1, x_2)^T : \frac{z-\mu}{\sigma} = \log(\frac{\alpha}{1-\alpha}), \text{ where } z = wx_1 + (1-w)x_2\}. \tag{4}$$

Let $\mathcal{X}$ be the pool of data. Suppose we have $(\boldsymbol{x}_1, Y_1), (\boldsymbol{x}_2, Y_2), \ldots, (\boldsymbol{x}_n, Y_n)$ in the training set. Based on this training data, we can estimate the threshold hyperplane $l_n = \{\boldsymbol{x} : F(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_n) = \alpha\}$ by

$$l_n : \quad \hat{w}_n x_1 + (1 - \hat{w}_n)x_2 = \hat{\mu}_n + \hat{\sigma}_n \log(\frac{\alpha}{1-\alpha}), \tag{5}$$

where $\hat{\boldsymbol{\theta}}_n = (\hat{\mu}_n, \hat{\sigma}_n, \hat{w}_n)^T$ is estimated from the labelled data $(\boldsymbol{x}_1, Y_1), (\boldsymbol{x}_2, Y_2), \ldots, (\boldsymbol{x}_n, Y_n)$. The details of the estimator $\hat{\boldsymbol{\theta}}_n$ are described in Section 3.2. Now, using the idea in stochastic approximation, we choose the next point from $\mathcal{X}$ as the closest to the estimated hyperplane. Note that we have to choose the closest point because none of

the points in $\mathcal{X}$ may fall on the hyperplane. Thus

$$\boldsymbol{x}_{n+1} = \arg\min_{\boldsymbol{x}\in\mathcal{X}} |F(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_n) - \alpha|. \tag{6}$$

There can be multiple points satisfying (6) because $\boldsymbol{x} \in \mathbb{R}^2$. Moreover, as pointed out in the previous section, the stochastic approximation method produces points clustered around the true hyperplane, which leads to poor estimation of some of the parameters in the model. We can overcome these problems by integrating the above approach with the optimal design approach.

First, we choose k_0 points as candidates which are closest to the estimated threshold hyperplane l_n. Denote them as $\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_{k_0}$. Then, we select the next point as the one maximizing the determinant of the Fisher information matrix among the candidates. Thus

$$\boldsymbol{x}_{n+1} = \arg\max_{\boldsymbol{x}\in\{\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{x}}_2,\ldots,\tilde{\boldsymbol{x}}_{k_0}\}} \det(I(\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n, \boldsymbol{x})). \tag{7}$$

The Fisher information matrix for $\boldsymbol{\theta}$ can be calculated as

$$I(\boldsymbol{\theta}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) = \sum_{i=1}^{n} \frac{e^{g(\boldsymbol{x}_i)}}{(1+e^{g(\boldsymbol{x}_i)})^2} \frac{\partial g(\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}} \frac{\partial g(\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}^T}, \tag{8}$$

where $g(\boldsymbol{x}) = (z-\mu)/\sigma$, $z = wx_1 + (1-w)x_2$ and $\boldsymbol{\theta} = (\mu, \sigma, w)^T$. The foregoing approach inherits the advantages of both stochastic approximation and optimal design. The stochastic approximation method in (6) can produce reasonable estimates of μ and σ, but can be very poor in the estimation of w. Because the D-optimality criterion in (7) ensures that the chosen points are well-spread, we can get a better estimate of w.

The improved estimation in our approach can be shown by considering the following version of the problem. Assume that there is at least one point in $\mathcal{X}$ that lies in the hyperplane l_n. Then, the selected point $\boldsymbol{x}_{n+1}$ is the solution of the following

optimization problem:

$$\max_{\boldsymbol{x}} \det(I(\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n, \boldsymbol{x})) \tag{9}$$
$$s.t.\ \hat{w}_n x_1 + (1 - \hat{w}_n) x_2 = \hat{\mu}_n + \hat{\sigma}_n \log(\frac{\alpha}{1-\alpha}).$$

As shown in the Appendix, it is equivalent to

$$\max_{\boldsymbol{x}} \boldsymbol{\eta}_x^T I^{-1}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) \boldsymbol{\eta}_x \tag{10}$$
$$s.t.\ \hat{w}_n x_1 + (1 - \hat{w}_n) x_2 = \hat{\mu}_n + \hat{\sigma}_n \log(\frac{\alpha}{1-\alpha}),$$

where $\boldsymbol{\eta}_x = (-1/\sigma, -\log(\alpha/(1-\alpha)), (x_1 - x_2)/\sigma)^T$. The objective function in (10) is precisely the estimated variance of the hyperplane where the data is collected. It gives us more accurate estimation when the response is acquired at the point with largest uncertainty. Thus, we select the point to be labelled such that the expected information gain is maximized. Note that the objective function in (10) is associated with $\boldsymbol{x}$ only through $\boldsymbol{\eta}_x$. It maximizes a quadratic form in terms of $(x_1 - x_2)$. Therefore, the optimal value is achieved on the boundary of the feasible region of $(x_1 - x_2)$. The point selected by (7) is expected to be not close to the previous selected ones when they are projected onto the estimated threshold hyperplane l_n. This is why the proposed approach can provide a more stable estimation of the parameter w. A pseudo code of the proposed approach is shown in Figure 3.4.1.

3.4.2 Estimation

Since (3) is a probabilistic model, it is tempting to consider maximum likelihood estimation (MLE) for the parameter $\boldsymbol{\theta}$. Suppose the labelled data are $(\boldsymbol{x}_1, Y_1)$, $(\boldsymbol{x}_2, Y_2)$, ..., $(\boldsymbol{x}_n, Y_n)$. It is known that the existence and uniqueness of MLE can be achieved only when successes and failures overlap (Silvapulle, 1981; Albert and Anderson, 1984; Santner and Duffy, 1986). However, even when we are able to compute the MLE, they may suffer from low accuracy due to the small sample size, especially for

Input: α value.

Suppose n data points are in the training set.
While Check stopping criterion,
Step 1, Find efficient estimators $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ from $(\boldsymbol{x}_i, Y_i)_1^n$.
Step 2, Choose k_0 candidate points which are closest to the estimated threshold hyperplane $\hat{l}_{\boldsymbol{x}} = \{\boldsymbol{x} : F(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_n) = \alpha\}$.
Step 3, Select the next point $\boldsymbol{x}_{n+1}$ by (7).
Step 4, Get the response Y_{n+1} for $\boldsymbol{x}_{n+1}$.
Step 5, Set n = n+1.
End

Output: $\hat{l}_{\boldsymbol{x}} = \{\boldsymbol{x} : F(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_n) = \alpha\}$.

Figure 11: The proposed active learning algorithm

nonlinear models. Use of a Bayesian approach with proper prior distribution for the parameters can overcome these problems.

We use the following priors:

$$\mu \sim \mathrm{N}(\mu_0, \sigma_\mu^2),\ \sigma \sim \text{Exponential}(\sigma_0),\ w \sim \text{Beta}(\alpha_0, \beta_0). \tag{11}$$

A normal prior is specified for the location parameter μ. The scale parameter σ is nonnegative since each x_i is assumed positively related with the response Y. Therefore, an exponential prior with mean σ_0 is used as the prior for σ. Because w is a weight factor in $[0, 1]$, a beta distribution is a reasonable prior for w.

Assuming μ, σ and w are independent with each other, the overall prior for $\boldsymbol{\theta}$ is the product of the priors for each of its components. Thus, the posterior distribution is

$$f(\boldsymbol{\theta}|\mathbf{Y}) \propto \prod_{i=1}^{n} \left(\frac{e^{(z_i-\mu)/\sigma}}{1+e^{(z_i-\mu)/\sigma}}\right)^{Y_i} \left(\frac{1}{1+e^{(z_i-\mu)/\sigma}}\right)^{1-Y_i} e^{\frac{(\mu-\mu_0)^2}{-2\sigma_\mu^2}} \lambda_0 e^{-\lambda_0\sigma} w^{\alpha_0-1}(1-w)^{\beta_0-1}, \tag{12}$$

where $z_i = wx_{i1} + (1-w)x_{i2}$ and $\boldsymbol{x}_i = (x_{i1}, x_{i2})^T$. Finding the posterior mean of the parameters is difficult because it involves a complicated multidimensional integration. The maximum-a-posterior (MAP) estimators are much easier to compute. The MAP

estimators of μ, σ, and w are obtained by solving

$$\hat{\boldsymbol{\theta}}_n = (\hat{\mu}_n, \hat{\sigma}_n, \hat{w}_n)^T = \arg\max_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}|\mathbf{Y}), \tag{13}$$

where

$$\begin{aligned}\log f(\boldsymbol{\theta}|\mathbf{Y}) \triangleq & \sum_{i=1}^{n} \frac{z_i - \mu}{\sigma} Y_i - \sum_{i=1}^{n} \log(1 + \exp(\frac{z_i - \mu}{\sigma})) \\ & - \frac{(\mu - \mu_0)^2}{2\sigma_\mu^2} - \lambda_0 \sigma + (\alpha_0 - 1)\log(w) + (\beta_0 - 1)\log(1 - w).\end{aligned} \tag{14}$$

Because proper prior distributions are employed, the optimization in (13) is well defined even when $n = 1$. Thus, this Bayesian approach allows us to implement a *fully* sequential procedure, i.e., the proposed active learning method can begin from $n = 1$. This would not have been possible with a frequentist approach (Wu, 1985), for which some initial sample is necessary before the active learning method can be called. One advantage of using initial sample is that the approach will be more robust to prior specifications. In Section 6, we report a simulation study to compare the use of initial sample against a fully sequential procedure.

3.5 Case Study

Financial institutions invest much resources and efforts into detection of money laundering. We applied the proposed method to some real transaction data from a financial institution. The data in this example consists of 92 accounts from personal customers. It keeps the recent two-year transaction history for each customer. By working with expert investigators, we got a large set of summary variables. Then using multi-stage modelling and dimension reduction on these summary variables, we extracted two statistical features $\boldsymbol{x} = (x_1, x_2)^T \in \mathbb{R}^2$. Based on discussions with expert investigators, these two features can be highly representative of the suspiciousness for the transaction history, where x_1 describes the velocity and amount of money flowing through the account, and x_2 measures the differences of the transaction behaviors among the peer comparisons. For reasons of confidentiality, we do not disclose

more details about the data being used here. Variables x_1 and x_2 are standardized to have zero mean and unit variance. The standardized data is shown in Figure 12.

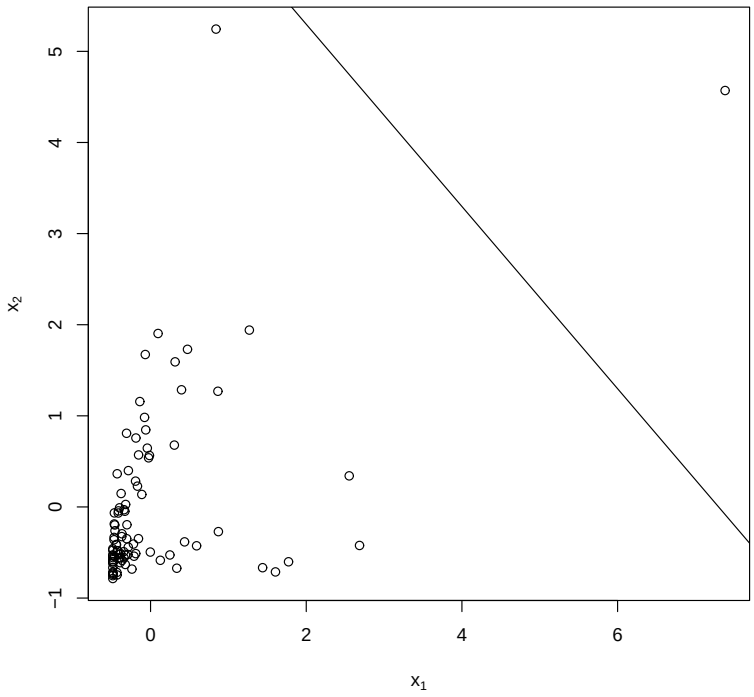


Figure 12: The standardized data. (Black line: the initial estimated threshold hyperplane by w_0, μ_0 and σ_0.)

We need to specify the prior for μ, σ, and w in (11) before active learning can be started. Here we use a heuristic procedure for doing this. First consider the prior for w. Assuming equal importance of x_1 and x_2 on the response, we would like the mean of w to be 0.5. Thus, set $\alpha_0/(\alpha_0 + \beta_0) = w_0 = 0.5$, which implies $\alpha_0 = \beta_0$. To get a flat prior, we take $\alpha_0 = \beta_0 = 3/2$. Thus, $w \propto w^{\frac{1}{2}}(1-w)^{\frac{1}{2}}$. Now consider the priors for μ and σ. Choose two extreme points (i.e., two accounts) $\boldsymbol{x}_l$ and $\boldsymbol{x}_u$ based on the lowest and highest values of z (denoted by z_l and z_u) through the mapping $z = w_0 x_1 + (1 - w_0)x_2$. We assume $\alpha_l = 5\%$ suspicious level for $\boldsymbol{x}_l$ and $\alpha_u = 95\%$

suspicious level for $\boldsymbol{x}_u$. Plugging them into the model (3), we obtain

$$z_l = \mu + \sigma \log \frac{\alpha_l}{1-\alpha_l},$$
$$z_u = \mu + \sigma \log \frac{\alpha_u}{1-\alpha_u}.$$

Now, μ_0 and σ_0 are obtained by solving the above equations as

$$\mu_0 = \frac{z_l \log \frac{\alpha_u}{1-\alpha_u} - z_u \log \frac{\alpha_l}{1-\alpha_l}}{\log \frac{\alpha_u}{1-\alpha_u} - \log \frac{\alpha_l}{1-\alpha_l}} = \frac{z_l + z_u}{2}, \tag{15}$$

$$\sigma_0 = \frac{z_u - z_l}{\log \frac{\alpha_u}{1-\alpha_u} - \log \frac{\alpha_l}{1-\alpha_l}}. \tag{16}$$

We take σ_μ^2 as the sample variance of z_i, $i = 1, \ldots, n$, where $z_i = w_0 x_{i1} + (1 - w_0) x_{i2}$. This completes the prior specification for the three parameters.

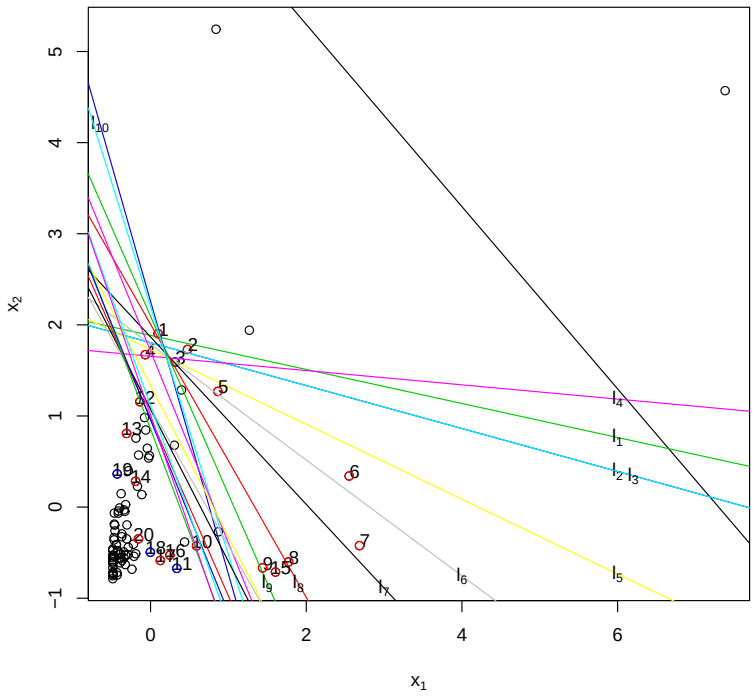


Figure 13: Active Learning via Sequential Design. (For example, yellow line l_5 stands for the estimated threshold hyperplane at iteration 5.)

Now the active learning method can be started. Suppose our objective is to find the threshold hyperplane with $\alpha = 0.75$. The initial estimated hyperplane based on

only the prior is shown in Figure 12. The points are then selected one at a time using the procedure described in the previous section. In this example, we took $k_0 = 15$ in (7). The performance of the proposed method for the first 20 points is shown in Figures 13 and 14.

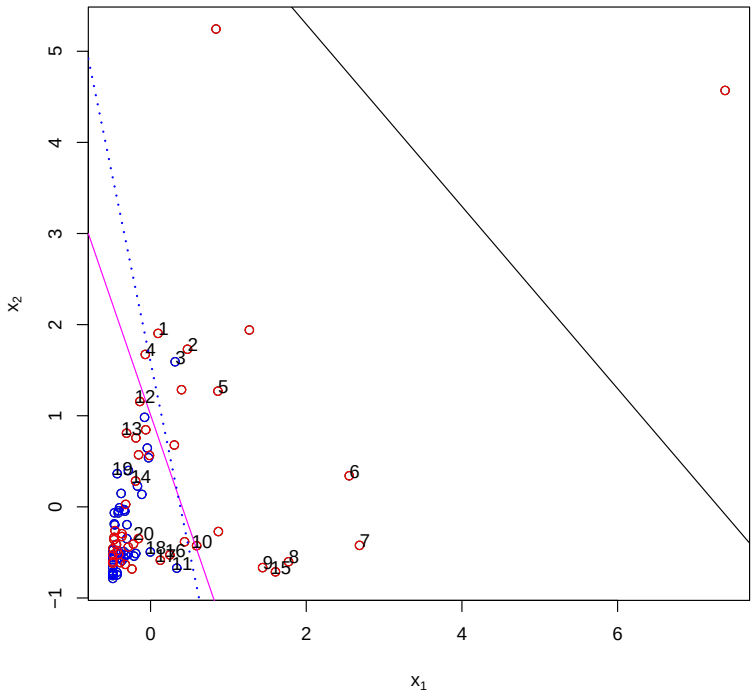


Figure 14: Comparison with the estimate based on full information. (Black line: the initial estimated threshold hyperplane by w_0, μ_0 and σ_0. Pink line: the estimated threshold hyperplane after 20 points are sequentially selected. Blue dashed line: the estimated threshold hyperplane when all data are labelled.)

Figure 13 shows a series of estimated threshold hyperplanes using the proposed approach. The red data point in the figure means it is selected and the response is 1. The blue one means it is selected and the response is 0. At the beginning, there were large changes in the threshold hyperplane. In about 10-15 points it started to converge. The final estimated threshold hyperplane (i.e., after 20 points) is shown in Figure 14. The points above this hyperplane should be given higher priority and be investigated thoroughly for their suspiciousness. There are only a few remaining

accounts that need a thorough investigation, which clearly shows the efficiency of the proposed method.

To assess the accuracy of the proposed method, we asked the investigators at this financial institution to investigate all the 92 accounts carefully. Based on the obtained information for all the accounts, we estimated the threshold hyperplane, which is shown in Figure 14 as blue dashed line. We can see that it is very close to the estimated threshold hyperplane (i.e., pink line) by the active learning method. Thus, the proposed method can identify the true hyperplane by using only about 22% ($\approx 20/92$) of the data, which is a big saving for the financial institution.

To check the efficiency of the proposed method, we also compared the proposed method with a naive method. The naive method is to randomly select the next data point for getting the response. To gauge the performance of two methods, we measure the closeness between the estimated threshold hyperplane l_n and the true threshold hyperplane $l_{\boldsymbol{x}}$ when all data are labelled. The adopted measure is

$$\text{dist}(l_n, l_{\boldsymbol{x}}) \triangleq \sum_{\mathbf{t}_i \in \mathbf{T}} d_i^2, \tag{17}$$

where $\mathbf{T} = \{\mathbf{t}_i\}$ is a set of points which lie on the true threshold hyperplane $l_{\boldsymbol{x}}$, and d_i is the distance of $\mathbf{t}_i$ to the estimated hyperplane l_n. Based on (17), a distance-based performance measure is defined as

$$Dist\_PM = \frac{1}{M} \sum_{j=1}^{M} \text{dist}_j(l_n, l_{\boldsymbol{x}}), \tag{18}$$

where M is the number of simulations, and dist_j represents $\text{dist}(l_n, l_{\boldsymbol{x}})$ for the j-th simulation.

Figure 15 shows the learning curves for the two methods. It is clear that the proposed method is much more efficient than the naive method. The estimated threshold hyperplane by the proposed method also moves towards the true threshold hyperplane quickly and consistently. The proposed method converges in about 10 steps for this problem.

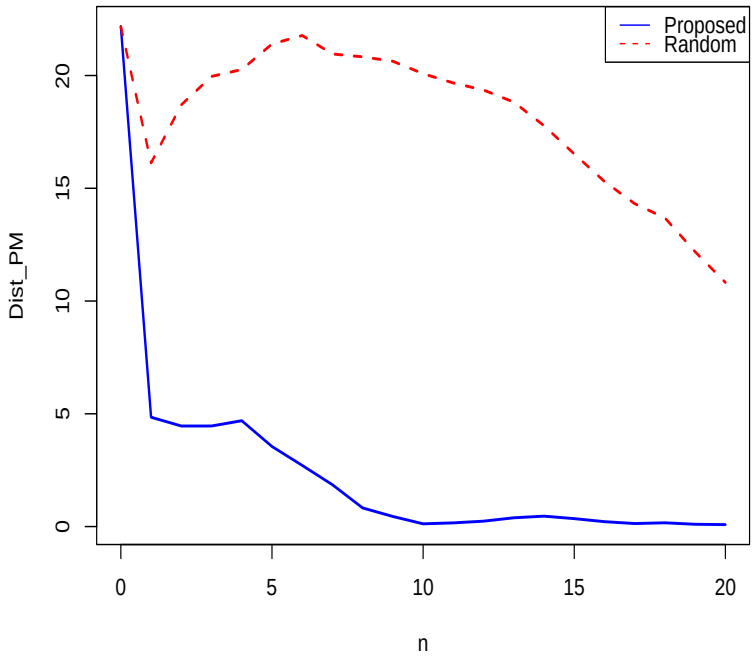


Figure 15: Learning Curves of Two Methods

3.6 Simulations

3.6.1 Numerical Examples

As stated before, the proposed method is expected to be more flexible and robust to model assumptions. Some experiments were conducted to study its performance. The simulated data were based on different models of $F(\boldsymbol{x})$. Four models were used in the study:

$$\text{Logistic distribution: } F(\boldsymbol{x}) = \frac{\exp(\frac{z-\mu}{\sigma})}{1+\exp(\frac{z-\mu}{\sigma})},$$

$$\text{Uniform distribution: } F(\boldsymbol{x}) = \frac{\frac{z-\mu}{\sigma} - (-2)}{2-(-2)},$$

$$\text{Normal distribution: } F(\boldsymbol{x}) = \Phi(\frac{z-\mu}{\sigma}),$$

$$\text{Cauchy distribution: } F(\boldsymbol{x}) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(\frac{z-\mu}{\sigma}),$$

where $z = wx_1 + (1-w)x_2$ and Φ is the standard normal distribution function. The true values of parameters were set as $\mu = 0.5$, $\sigma = 1$ and $w = 0.7$. The response outcome at each point was generated according to $F(x)$.

In this simulation we chose $\alpha = 0.5$ and $\alpha = 0.8$ for illustration. The same performance measure in (18) is used here. Let $k_0 = 15$ in (7). The specification of hyper-parameters is done by using the heuristic procedure discussed in the previous section. 100 simulations were performed and $n = 30$ points were sequentially selected in each simulation.

Several methods are considered for comparison. We denote the fully sequential version of the proposed method as Method I. We denote by Method II the proposed method whose iterative scheme is preceded by choosing a fixed initial sample. To get a baseline comparison we used Method III, where the points are selected randomly, i.e., without using any active learning method. For Method II, we used stratified random sampling to choose eight initial points. It is implemented as follows. With the initial guess on the parameters μ_0, σ_0 and w_0, we can get $z = w_0x_1 + (1-w_0)x_2$. Then we divide the range of z into four strata as $(-\infty, \mu_0 - 1.6\sigma_0)$, $[\mu_0 - 1.6\sigma_0, \mu_0)$, $[\mu_0, \mu_0 + 1.6\sigma_0)$ and $[\mu_0 + 1.6\sigma_0, +\infty)$. Since each point $\boldsymbol{x}$ can be mapped into the z value, we randomly choose two $\boldsymbol{x}'s$ in each stratum according to the corresponding z value. The choice of the constant ± 1.6 is based on the asymptotic optimality of the estimators under logistic distribution (see, e.g., Neyer 1994). Performance of the three methods for two chosen values of α are shown in Figures 16 and 17.

Clearly the proposed active learning methods (I and II) perform much better than Method III. Between I and II, Method I outperforms Method II. This is expected because Method I starts the active learning from the first point, whereas Method II starts active learning only after the selection of eight initial points. Unless stratified samples can be properly chosen, Method II will give inferior results. However, as n increases to about 30, its performance is comparable to that of Method I.

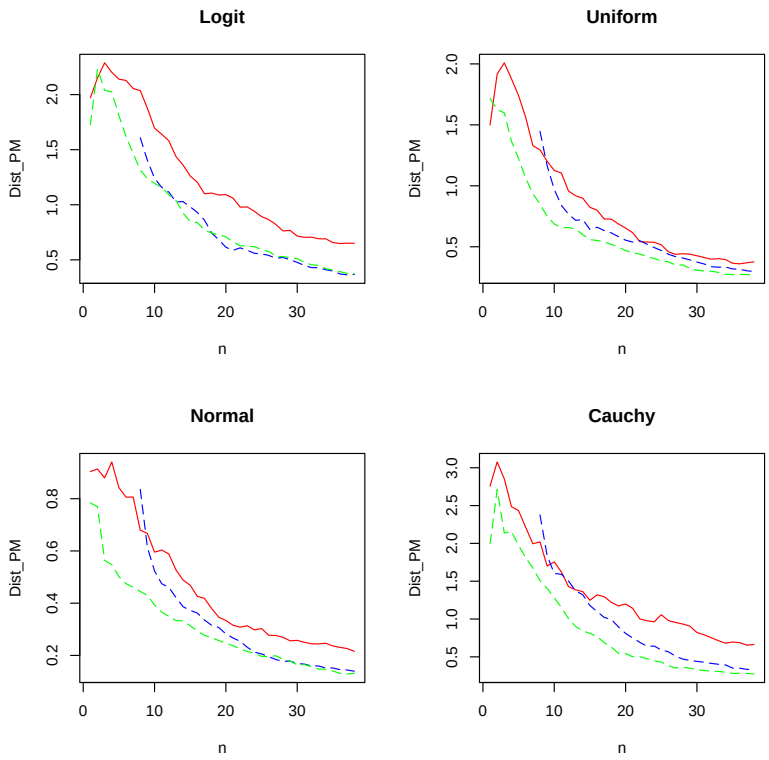


Figure 16: Dist_PM for four models with $\alpha = 0.5$. (Green line: method I. Blue line: method II. Red line: method III.)

Comparing Figures 16 and 17, we can see that the performance of the methods is better when $\alpha = .5$. This is a well known fact in the literature that the estimation of extreme quantiles is much more difficult than with $\alpha = .5$ (see, e.g., Joseph 2004). It is also clear from the figures that the proposed methods are quite robust to model assumptions.

In the proposed active learning approach in (7), one selects k_0 candidate points which are closest to the estimated hyperplane. Here k_0 is considered as a tuning parameter but its optimal value has not yet been addressed. An additional experiment was conducted regarding the choice of k_0. Setting $\alpha = 0.6$, the proposed active learning in a fully sequential version (i.e., Method I) is performed for different k_0, i.e., $k_0 = 1$, 5, 10, 15 and $k_0 = N$, where N is the total number of data points in the data set. $k_0 = 1$ means active learning using stochastic approximation, whereas

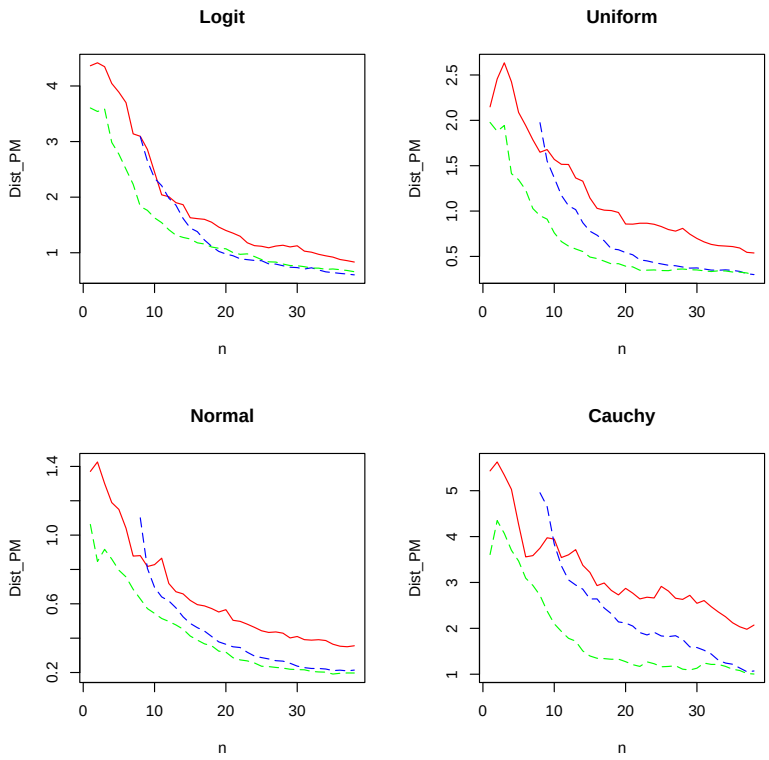


Figure 17: Dist_PM for four models with $\alpha = 0.8$. (Green line: method I. Blue line: method II. Red line: method III.)

$k_0 = N$ means active learning using a fully D-optimal-based sequential design. 100 simulations were generated for each k_0 and each model. The hyper-parameters were chosen as in Section 4. Figure 18 shows the simulation results .

As can be seen in Figure 18, except for the logistic distribution the Dist_PM decreases up to some value of k_0 and then increases. This agrees with our initial intuition that choosing a large value of k_0 may not be good if the assumed model is not correct. Our procedure assumes the logistic model. Thus, when the model is changed to uniform, normal, or Cauchy, the method did not do well with a large k_0. As expected, the performance did not deteriorate with k_0 when the true model is logistic. It is also clear that $k_0 = 1$ is a bad choice as the Dist_PM is the largest in all cases. Thus, using a purely stochastic approximation method for active learning is not good in this particular problem. It is not clear what is the best value of k_0.

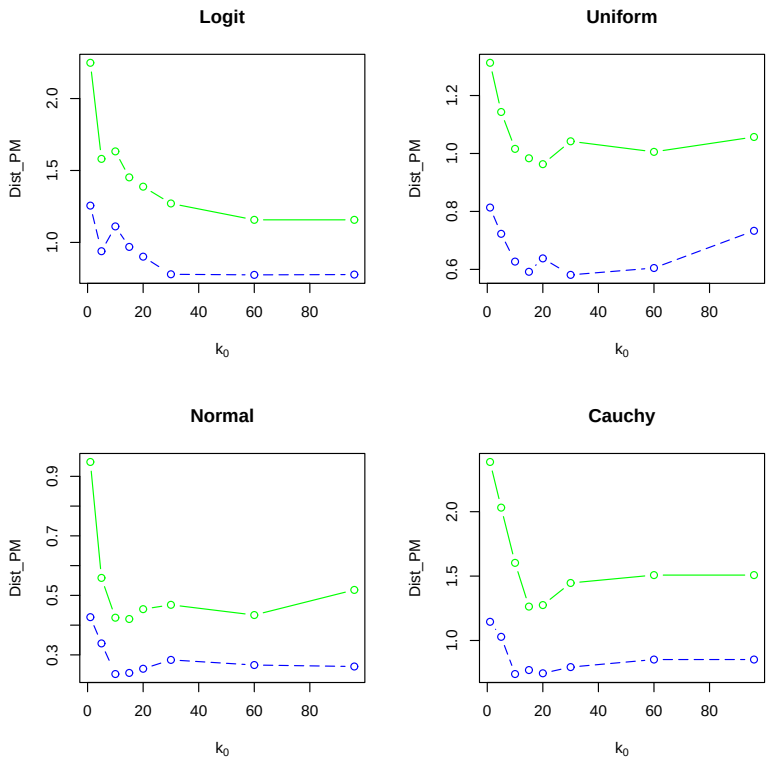


Figure 18: Performance with different k_0. (Green line: $n = 10$; Blue line: $n = 20$.)

The simulation results suggest choosing k_0 to be 20%-50% of N.

3.6.2 Comparison with Support Vector Machine

Active learning using support vector machine (SVM) for classification has been proposed with several versions (Schohn & Cohn, 2000; Campbell et al., 2000; Tong & Koller, 2001). The basic idea is to label points that lie closest to the SVM's dividing hyperplane. It is known that the hyperplane in SVM converges to the Bayes rule $P(Y = 1|\boldsymbol{x}) = \alpha$, where $\alpha = 0.5$. The proposed active learning via sequential design can also converge to the threshold hyperplane when $\alpha = 0.5$. To start the active learning with SVM, some initial sample of points are needed. Therefore, to have a fair comparison, we used eight points as the initial sample chosen based on the stratified random sampling discussed in Section 5.1. The hyper-parameters were chosen as before. 100 simulations were generated for comparison.

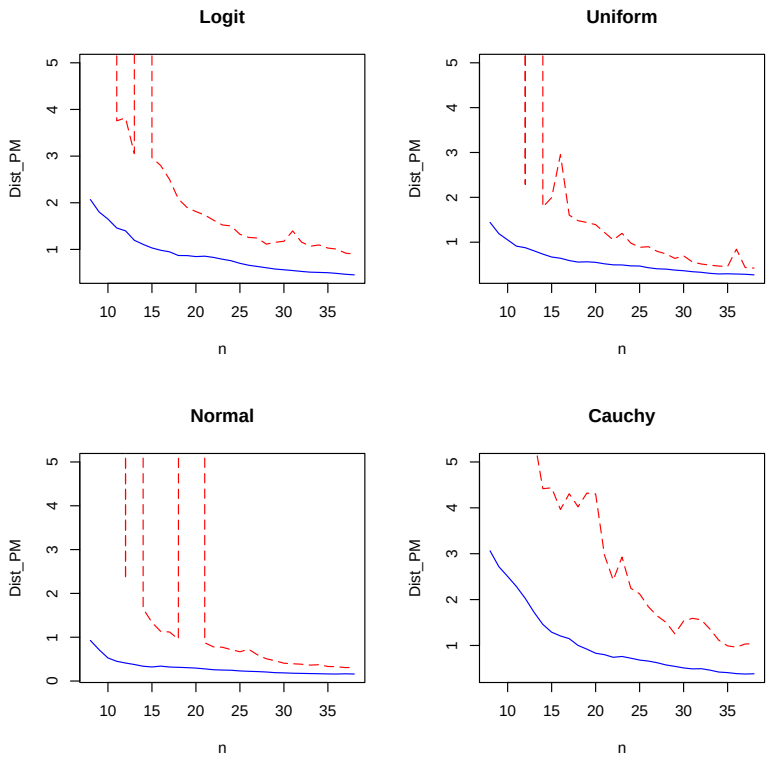


Figure 19: Comparison of Active Learning with SVM. Solid blue line: the proposed active learning (method II). Dashed red line: active learning with SVM.

The Dist_PM values are plotted in Figure 19. We can see that the Dist_PM values of the proposed active learning (Method II) are much smaller than that of the active learning with SVM. Moreover, the proposed active learning is quite stable, whereas the SVM is quite unstable for small n. The SVM is not robust because adding one more point into the training set can cause big changes in the SVM's dividing hyperplane. Thanks to the use of the Bayesian approach, the estimation in the proposed active learning is stable.

The proposed active learning seems to converge within 20 steps, while the active learning with SVM needs at least 10 more steps to achieve similar performance. The imporvement is even more pronounced with heavy tail distributions like Cauchy. Thus in this particular problem, the proposed active learning outperformed active learning with SVM in all aspects including accuracy, stability, and robustness.

3.7 Discussions and Conclusions

In this chapter, we propose an active learning via sequential design and report its application to a real world problem in money laundering detection. Due to the large amount of transactions and various business categories, it is crucial to find an efficient way to get the threshold hyperplane for prioritization. The proposed method is efficient and accurate for estimating the threshold hyperplane, and its performance is robust to model assumptions. It can help investigation to put more effort on those accounts with great importance. Therefore, this approach can significantly improve the productivity of money laundering detection.

The propose active learning method uses a combination of stochastic approximation and optimal design methods. From the sequential design perspective, we have shown that the proposed method works better than both stochastic approximation and optimal design. Through simulations we have also shown that the proposed method outperforms active learning methods using SVM. With proper prior information, the fully sequential version of the proposed method (Method I) performs better than the one which starts with an initial sample (Method II). Regarding the choice of k_0 (i.e., the number of candidate points in (7)), the simulation study suggests choosing k_0 to be 20%-50% of N.

The proposed method is described for two variables $\boldsymbol{x} = (x_1, x_2)^T$. It can be easily extended to high dimensions. In multivariate situations where $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$, we can define a synthetic variable z as a convex combination of the feature variables, i.e., $z = \sum_{i=1}^{p} w_i x_i$, where $w_i \geq 0$ and $\sum_{i=1}^{p} w_i = 1$. Then the active learning procedure is the same as the one described in Section 4.1. The prior for $w = (w_1, w_2, \ldots, w_p)^T$ can be chosen to be a Dirichlet distribution.

The proposed active learning via sequential design is flexible in estimating threshold hyperplane for different α. On the other hand, the standard support vector machine is mainly for classification problem with $\alpha = 0.5$. Lin et al. (2002) proposed

a modified support vector machine to account for α different from 0.5. It will be interesting to compare the proposed method with active learning using the modified support vector machine. Note, however, the absence of active learning method using the modified support vector machine in the literature.

Although the proposed method was motivated by the problem of detecting money laundering, the approach is quite general and can be applied to different classes of problems. For example, it can be used in sensitivity experiments (Neyer, 1994) or in bioassay experiments (McLeish and Tosh, 1990). Another advantage of the proposed method is that it can be applied to multivariate problems. For this to work, we need to assume the direction of the effect for each of the variables. This assumption seems to be reasonable in problems we have encountered so far.

CHAPTER IV

FACTOR LOGIT-MODELS WITH A LARGE NUMBER OF CATEGORIES

4.1 Introduction

Multi-category classification has been drawn great attention both in machine learning and statistics community. It has been widely used into many fields such as microarray gene expression, pattern recognition, etc. Among many proposed multi-category classification methods in literature, there are mainly two ways to tackle the problem. One is to solve the problem by handling a series of binary classifications (Dietterich and Bakiri, 1995; Schapire, 1997; Allwein et al., 2000). Although solving a series of binary classifications is popular in multi-category classification methods, this pairwise approach has the disadvantage of potential variance increase since smaller samples are utilized to learn each classifier function. The other is to consider all classes at once and estimate all classifier functions from one loss function. For example, the multi-logit model, a generalization of binary logit, is one of the most common methods to learn all classes at once.

To get a better insight of multi-category classification methods, study of theoretical properties of the estimated classifier functions is of great importance. There have been some work to study the consistency of classifier functions for multi-category classification methods (Zhang, 2004; Tewari and Bartlett, 2007). Intuitively, the convergence rate of estimates will slow down as the number of categories increases. However, the exact relations between convergence rates and number of categories are not that obvious. Compared with binary classification problem, the multi-category classification method has to estimate a series of classifier functions simultaneously.

The classifier functions of different categories can interact with each other, which brings more challenges to study their asymptotic behaviors.

It is worth noting that when the number of categories is relatively large, the classifier functions are likely to be located in a functional subspace with much smaller dimensions than the number of categories. Motivated by this observation, we propose a factor model for classifier functions under a penalized multi-logit model. It means that each classifier function can be determined by several common factors. In this situation, we show that the convergence rate of the classifier functions is not relied on the number of categories, but only dependent on the number of factors under the optimal tuning parameter. Therefore, when the number of factors is much smaller than the number of categories, the proposed method can achieve better convergence rate of classifier functions.

The remaining of the chapter is organized as follows. The framework and main results of this work is described in Section 4.2. The asymptotic properties of the proposed factor logit-models are studied in Section 4.3. The performance of the proposed method is illustrated using simulation in Section 4.4. Some discussions and future works are concluded in Section 4.5. The technical proofs are put in the appendix.

4.2 Framework and Main Results

Without loss of generality, let us assume the class label $Y \in \{1, 2, \ldots, K\}$, where K is the number of classes. Suppose $Y \sim$ Multi-nomial$(p_1, \ldots, p_K)$, where $\sum_{i=1}^{K} p_k = 1$. The multi-logit model is

$$p_k(\boldsymbol{f}(\boldsymbol{x})) \triangleq P(Y = k|X = \boldsymbol{x}) = \frac{e^{f_k(\boldsymbol{x})}}{1 + \sum_{j=1}^{K-1} e^{f_j(\boldsymbol{x})}}, \tag{19}$$

where $\boldsymbol{X} \in \mathbb{R}^d$. The classifier function vector $\boldsymbol{f}$ is coded with a baseline constraint, i.e., $\boldsymbol{f} = (f_1, \ldots, f_{K-1}, 0)$. Suppose $p_k(\boldsymbol{f}(\boldsymbol{x}))$ is bounded away from zero and one.

Estimating the function vector $\boldsymbol{f}$ is the main of interest. The expectation of log-likelihood function based on (19) is

$$\begin{aligned} E(\log L(Y, \boldsymbol{f})) &= \sum_{k=1}^{K} P(Y = k|X) \log \frac{e^{f_k(x)}}{1 + \sum_{j=1}^{K-1} e^{f_j(x)}} \\ &= \sum_{k=1}^{K} P(Y = k|X)[f_k(x) - \log(1 + \sum_{j=1}^{K-1} e^{f_j(x)})]. \end{aligned} \tag{20}$$

In order to get good and smooth estimate of classifier functions, we consider to use regularization as a tool to penalize the roughness of the estimates. Suppose a penalty function is $J(\boldsymbol{f})$, the generalized risk function with penalty can be defined as

$$l_\lambda(\boldsymbol{f}) = R(Y, \boldsymbol{f}) + \lambda J(\boldsymbol{f}), \qquad \lambda > 0. \tag{21}$$

Here $R(Y, \boldsymbol{f}) = -2E(\log L(Y, \boldsymbol{f}))$ and λ is the regularization parameter. The penalty functional J is assumed to have a form $J(\boldsymbol{f}) = \sum_{i=1}^{K-1} \langle f_i, w_i f_i \rangle$, where each w_i is positive definite.

Since $X = (x^{(1)}, \ldots, x^{(d)})^T \in \mathbb{R}^d$, each classifier function f_i is a multivariate function. One major difficulty in estimation is caused by the fact of curse of dimensionality. The convergence rate of the estimated classifier functions can be much slower in high dimensional problems than those in low dimensional problem. To bypass the difficulty, one popular choice is to use the additive models (Stone, 1985; Hastie and Tibshirani, 1990). Generally, additive models assume that a multivariate function can be decomposed as a sum of one dimensional functions. Adopting this strategy, we decompose each classifier function $f_i(x)$ as

$$f_i(x) = \mu + h_{i1}(x^{(1)}) + \cdots + h_{id}(x^{(d)}), \tag{22}$$

where μ is a constant function and the identifiability of each component in (22) is assured by the side conditions. Here each $h_{ij} \in \mathcal{H}_j$, where $\mathcal{H}_j$ is a functional space of one dimensional function on $x^{(j)}$. Stone (1985) showed that the optimal convergence rate for the additive models is the same as that for the univariate function estimation

problem. Let $\mathcal{H}_j = \{1\} \oplus \bar{\mathcal{H}}_j$, then we can define the full functional space $\mathcal{H}$ for classifier functions as

$$\mathcal{H} = \{1\} \oplus \bar{\mathcal{H}}_1 \oplus \cdots \oplus \bar{\mathcal{H}}_d. \tag{23}$$

We also assume that each functional space $\mathcal{H}_j$ on x_j is a Soblev space, which is given by

$$W_2^m(\Omega) = \{h : \Omega \to R | h, h^1, \ldots, h^{m-1} \text{ are absolutely continuous and } h^m \in L_2(\Omega)\},$$

where Ω is the domain for $x^{(j)}$. This is a natural choice given the penalty. There are many possible inner products on W_2^m under which it is a Hilbert space.

Suppose the observations are (x_i, y_i), $i = 1, \ldots, n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{1, 2, \ldots, K\}$. Let us code $y_i = (y_{i1}, y_{i2}, \ldots, y_{iK})$ where $y_{ik} \in \{0, 1\}$ and for any k, y_{ik} is equal to 1 if $y_i = k$, 0 otherwise. Now the empirical risk function with the penalty term can be written as

$$\begin{aligned} l_{n\lambda}(\boldsymbol{f}) &= -2\left[\frac{1}{n}\log\prod_{i=1}^{n}\prod_{k=1}^{K}(P_k(\boldsymbol{f}(x)))^{y_{ik}}\right] + \lambda J(\boldsymbol{f}) \\ &= -2\left\{\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[y_{ik}(f_k(x_i) - \log(1 + \sum_{j=1}^{K-1} e^{f_j(x_i)}))\right]\right\} + \lambda J(\boldsymbol{f}) \\ &\triangleq l_n(\boldsymbol{f}) + \lambda J(\boldsymbol{f}). \end{aligned} \tag{24}$$

The estimates of classifier function vector $\boldsymbol{f}$ is obtained by minimizing $l_{n\lambda}(\boldsymbol{f})$. For convenience, define $l(\boldsymbol{f}) = R(Y, \boldsymbol{f})$, which represents the generalized risk without the penalty term. Some notation are used here,

$$\boldsymbol{f}^0 = \arg\min_{\boldsymbol{f}\in\mathcal{H}} l(\boldsymbol{f}) \Rightarrow Dl(\boldsymbol{f}^0) = 0, \tag{25}$$

$$\tilde{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}\in\mathcal{H}} l_\lambda(\boldsymbol{f}) \Rightarrow Dl_\lambda(\tilde{\boldsymbol{f}}) = 0, \tag{26}$$

$$\hat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}\in\mathcal{H}} l_{n\lambda}(\boldsymbol{f}) \Rightarrow Dl_{n\lambda}(\hat{\boldsymbol{f}}) = 0, \tag{27}$$

where D is the notation for Frechet derivative (Rudin, 1991). For large n, with probability one, the risk functions $l_\lambda(\boldsymbol{f})$ and $l_{n\lambda}(\boldsymbol{f})$ are convex forms of $\boldsymbol{f}$. Therefore, $\tilde{\boldsymbol{f}}$ and $\hat{\boldsymbol{f}}$ defined in (26) and (27) can be uniquely determined respectively. Obviously, it can be seen that $\boldsymbol{f}^0$ follows the Bayes rule since

$$\begin{aligned} & Dl(\boldsymbol{f}^0) = 0, \\ \Rightarrow & \frac{\partial}{\partial f_k} R(Y, \boldsymbol{f}) = -2[P(Y = k|X) - \frac{e^{f_k(x)}}{1 + \sum_{j=1}^{K-1} e^{f_j(x)}}] = 0, \\ \Rightarrow & P(Y = k|X) = \frac{e^{f_k^0(x)}}{1 + \sum_{j=1}^{K-1} e^{f_j^0(x)}}. \end{aligned} \tag{28}$$

It is known that the convergence rate of classifier functions is associated with the number of categories K. It means that the estimate efficiency can deteriorate as the number of categories increases. However, often times and in many applications with a large number of categories, the relations among the classifier functions can be modelled in a lower dimensional functional space. In this situation, the convergence property of each classifier function will depend on the lower dimensional model. Therefore, we can use a factor model for the classifier functions.

Recall the multi-logit model in (19), suppose that each classifier function f_k can be described by a linear combination of a set of functions $g_1, g_2, \ldots, g_L$, i.e.,

$$f_k(x) = \sum_{r=1}^{L} \beta_{kr} g_r(x), \ k = 1, \ldots, K-1, \tag{29}$$

where β_{kr}'s are the parameters, and g_r is called the basis function. Note that a lower dimensional ($L << K$) model (29) is highly desirable. Statistical analysis can benefit from this model since it provides a large reduction of dimensionality. Furthermore, $g_1, \ldots, g_L$ can lead to insight into common properties of $f_1, \ldots, f_{K-1}$. With the prior knowledge of L, we use the data themselves to estimate appropriate basis functions.

With the observed data $x_1, \ldots, x_n$, we define F as

$$F \triangleq \begin{pmatrix} f_1(x_1) & \ldots & f_{K-1}(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_n) & \ldots & f_{K-1}(x_n) \end{pmatrix} = \begin{pmatrix} g_1(x_1) & \ldots & g_L(x_1) \\ \vdots & \ddots & \vdots \\ g_1(x_n) & \ldots & g_L(x_n) \end{pmatrix} \begin{pmatrix} \beta_{11} & \ldots & \beta_{K-1,1} \\ \vdots & \ddots & \vdots \\ \beta_{1L} & \ldots & \beta_{K-1,L} \end{pmatrix}. \tag{30}$$

Denoting $\tilde{\beta}_{ij} = \beta_{ij} / \sqrt{\sum_{i=1}^{K-1} \beta_{ij}^2}$, the matrix F can be rewritten as

$$\begin{aligned} F &= \begin{pmatrix} g_1(x_1) & \ldots & g_L(x_1) \\ \vdots & \ddots & \vdots \\ g_1(x_n) & \ldots & g_L(x_n) \end{pmatrix} \begin{pmatrix} \sqrt{\sum_{j=1}^{K-1} \beta_{j1}^2} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sqrt{\sum_{j=1}^{K-1} \beta_{jL}^2} \end{pmatrix} \begin{pmatrix} \tilde{\beta}_{11} & \ldots & \tilde{\beta}_{K-1,1} \\ \vdots & \ddots & \vdots \\ \tilde{\beta}_{1L} & \ldots & \tilde{\beta}_{K-1,L} \end{pmatrix} \\ &\triangleq GD^{1/2}B^T. \end{aligned}$$

Note that the above decomposition of F provides a way to generate the basis function g_r. Let us define the basis vector as $\vec{g}_r = (g_r(x_1), \ldots, g_r(x_n))^T$ where $1 \leq r \leq L$. To make the factor model identifiable, we impose suitable normalizing conditions, which are stated as follows:

(i) $(\frac{1}{\sqrt{n}}\vec{g}_r)^T \frac{1}{\sqrt{n}}\vec{g}_r = \delta_{rs}$;

(ii) $\sum_{j=1}^{K-1} \tilde{\beta}_{jr}\tilde{\beta}_{js} = 0$ if $r \neq s$;

(iii) $\sum_{j=1}^{K-1} \beta_{j1}^2 \geq \sum_{j=1}^{K-1} \beta_{j2}^2 \geq \cdots \sum_{j=1}^{K-1} \geq \beta_{jL}^2$.

Here $\delta_{rs} = 1$ if $r = s$, and $\delta_{rs} = 0$ otherwise. Obviously, the normalization given by (i)-(iii) depends on n, K and the observed data. As $n \to \infty$, it is easy to see that Condition (i) is asymptotically equivalent to choosing orthonormal functions (with respect to the L_2 norm). Condition (iii) provides an ordering on the basis functions. Considering the existence and uniqueness of basis functions and parameters satisfying

condition (i)-(iii), model (29) implies that

$$M = \frac{1}{n}FF^T = \frac{1}{n}GD^{1/2}B^TBD^{1/2}G^T$$
$$= \sum_{r=1}^{L}\left(\sum_{j=1}^{K-1}\beta_{j1}^2\right)\frac{1}{\sqrt{n}}\vec{g}_r\frac{1}{\sqrt{n}}\vec{g}_r^T.$$

It follows that $\sum_{j=1}^{K-1}\beta_{j1}^2, \sum_{j=1}^{K-1}\beta_{j2}^2, \ldots$ are the largest, second largest, $\ldots$ eigenvalues of M, and that $\frac{1}{\sqrt{n}}\vec{g}_1, \frac{1}{\sqrt{n}}\vec{g}_2, \ldots$ are the corresponding orthonormal eigenvectors. Therefore $\vec{g}_r$, $1 \le r \le L$, can be uniquely determined up to sign changes.

Let us denote $\vec{f}(x) = (f_1(x), \ldots, f_{K-1}(x))^T$. Based on the estimate of $\vec{f}(x)$, we can obtain estimate of basis functions from the eigenvectors of matrix $\hat{M}$. Here $\hat{M}$ is derived from $\hat{F}$ in the same fashion as F in (30), i.e.,

$$\hat{M} = \frac{1}{n}\hat{F}\hat{F}^T = \frac{1}{n}\begin{pmatrix} \vec{\hat{f}}(x_1)^T\vec{\hat{f}}(x_1) & \ldots & \vec{\hat{f}}(x_1)\vec{\hat{f}}(x_n) \\ \vdots & \ddots & \vdots \\ \vec{\hat{f}}^T(x_n)\vec{\hat{f}}(x_1) & \ldots & \vec{\hat{f}}^T(x_n)\vec{\hat{f}}(x_n) \end{pmatrix},$$

where $\vec{\hat{f}}(x) = (\hat{f}_1(x), \ldots, \hat{f}_{K-1}(x))^T$. Recall $\boldsymbol{f}^0, \tilde{\boldsymbol{f}}$, and $\hat{\boldsymbol{f}}$ defined in (25)-(27), we can define $\tilde{M}$ and M^0 in a similar fashion using $\tilde{\boldsymbol{f}}$ and $\boldsymbol{f}^0$ respectively. Note that the main of interest is to investigate convergence property of $\hat{\boldsymbol{f}}$ to $\boldsymbol{f}^0$. Under the factor model (29), the convergence properties of $\hat{\boldsymbol{f}}$ to $\boldsymbol{f}^0$ will depend on the convergence for the eigenvector $\vec{\hat{g}}_r$ of $\hat{M}$ to the eigenvector $\vec{g}_r^0$ of M^0. We study the asymptotic properties of convergence rate on $\|\vec{\hat{g}}_r - \vec{g}_r^0\|_2$, where $\|\cdot\|_2$ is the usual L_2 norm. It is shown that the optimal convergence rate is not dependent on the number of categories K, but relied on the number of factors L in (29).

Theorem 1. *Suppose the factor model in* (29) *is valid for* $\boldsymbol{f}^0 \in \mathcal{H}$. *Assume that* $p_i \sim 1/K$, $i = 1, \cdots, K$. *If* λ_n *is a sequence of positive numbers such that* $\lambda_n \to 0$, *then for any* $r \in \{1, \ldots, L\}$,

$$\frac{1}{n}\|\vec{\hat{g}}_r - \vec{g}_r^0\|_2^2 \le C_r[\lambda_n K + (\frac{n}{L})^{-1}(\lambda_n K)^{\frac{1}{2m}}],$$

where C_r is a constant independent of λ_n, n, and L. Here L is the number of factor in (29). Furthermore, let $\lambda_n = \frac{1}{K}(\frac{n}{L})^{-\frac{2m}{2m+1}}$, the resulting rate of convergence of the penalized likelihood estimator is

$$\frac{\sum_{k=1}^{K} \|\hat{f}_k - f_k^0\|_2^2}{K} \leq C(\frac{n}{L})^{-\frac{2m}{2m+1}} \quad w.p.\ 1,$$

where C is a constant independent of λ_n, n, and L.

4.3 Factor Multi-Logit Model

In this section, we study the asymptotic properties of the proposed factor multi-logit model for classifier functions. The proof of Theorem 1 is described in detail in Section 4.3.2 and 4.3.3. The general structure of the proof is decomposed into two parts. Obviously,

$$\vec{\hat{g}}_r - \vec{g}_r^0 = (\vec{\hat{g}}_r - \vec{\tilde{g}}_r) + (\vec{\tilde{g}}_r - \vec{g}_r^0), \tag{31}$$

where $\vec{\tilde{g}}_r$ is the eigenvector from $\tilde{M}$. Using the triangle inequality, we have

$$\|\vec{\hat{g}}_r - \vec{g}_r^0\|_2 \leq \|\vec{\hat{g}}_r - \vec{\tilde{g}}_r\|_2 + \|\vec{\tilde{g}}_r - \vec{g}_r^0\|_2. \tag{32}$$

We start to develop some theoretical properties for the estimated classifier functions of the penalized multi-logit model. Clearly,

$$\hat{\boldsymbol{f}} - \boldsymbol{f}^0 = (\tilde{\boldsymbol{f}} - \boldsymbol{f}^0) + (\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}).$$

We denote the term $\tilde{\boldsymbol{f}} - \boldsymbol{f}^0$ as systematic error and $\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}$ as stochastic error. Then the upper bound of $\|\vec{\hat{g}}_r - \vec{g}_r^0\|_2$ is developed in Section 4.3.2 based on $\tilde{\boldsymbol{f}} - \boldsymbol{f}^0$. Correspondingly, The upper bound of $\|\vec{\hat{g}}_r - \vec{\tilde{g}}_r\|_2$ are studied in Section 4.3.3 based on $\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}$.

4.3.1 Multi-Logit Model

In this section, we come to study the penalized multi-logit model. We divide $\hat{\boldsymbol{f}} - \boldsymbol{f}^0$ into two parts as the systematic error $\tilde{\boldsymbol{f}} - \boldsymbol{f}^0$ and stochastic error $\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}$. This

approach to study the property of penalized likelihood estimates has been used by many authors such as Silverman (1982), Cox and O'Sullivan (1990), and Lin (2000).

4.3.1.1 Systematic Error

Let us define $Z_\lambda(\boldsymbol{f}) \triangleq Dl_\lambda(\boldsymbol{f})$ and $G_\lambda(\boldsymbol{f}) \triangleq D^2 l_\lambda(\boldsymbol{f})$, where D is Frechet derivative. Then

$$Z_\lambda(\boldsymbol{f}) = D \cdot R(L(Y, \boldsymbol{f})) + \lambda W \boldsymbol{f},$$

where $W = \text{diag}(w_1, w_2, \ldots, w_{K-1})$, and

$$G_\lambda(\boldsymbol{f}) = D^2 \cdot R(L(Y, \boldsymbol{f})) + \lambda W.$$

From (25) and (26), we know that $Z_\lambda(\tilde{\boldsymbol{f}}) = 0$ and $D \cdot R(L(Y, \boldsymbol{f}^0)) = 0$. Linearizing $Z_\lambda(\boldsymbol{f})$ at $\boldsymbol{f}^0$ by Taylor expansion, it gives

$$Z_\lambda(\tilde{\boldsymbol{f}}) - Z_\lambda(\boldsymbol{f}^0) \approx (\tilde{\boldsymbol{f}} - \boldsymbol{f}^0)^T G_\lambda(\boldsymbol{f}^0). \tag{33}$$

Hence,

$$\begin{aligned} \tilde{\boldsymbol{f}} - \boldsymbol{f}^0 &\approx -G_\lambda^{-1}(\boldsymbol{f}^0) Z_\lambda(\boldsymbol{f}^0) \\ &= -G_\lambda^{-1}(\boldsymbol{f}^0)(\lambda W \boldsymbol{f}^0). \end{aligned} \tag{34}$$

In order to get a close form of $G_\lambda^{-1}(\boldsymbol{f}^0)$, we need to know the expression of $G_\lambda(\boldsymbol{f}^0)$. It requires the second derivatives of $l_\lambda(\boldsymbol{f}^0)$, which are

$$\begin{aligned} \frac{\partial^2 l_\lambda(\boldsymbol{f}^0)}{\partial^2 f_k} &= 2\left[\frac{e^{f_k^0(x)}[1 + \sum_{i \neq k}^{K-1} e^{f_i^0(x)}]}{(1 + \sum_{j=1}^{K-1} e^{f_j^0(x)})^2} + \lambda w_k\right] \\ &= 2[P(Y = k|X)(1 - P(Y = k|X)) + \lambda w_k], \end{aligned} \tag{35}$$

and if $k \neq m$,

$$\begin{aligned} \frac{\partial^2 l_\lambda(\boldsymbol{f}^0)}{\partial f_k \partial f_m} &= -2\frac{e^{f_k^0(x)} e^{f_m^0(x)}}{(1 + \sum_{j=1}^{K-1} e^{f_j^0(x)})^2} \\ &= -2P(Y = k|X)(1 - P(Y = m|X)). \end{aligned} \tag{36}$$

For simplicity, ignoring the constant 2 in (35) and (36), so

$$G_\lambda(\boldsymbol{f}^0) = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_{K-1} \\ -p_1p_2 & p_2(1-p_2) & \dots & -p_2p_{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{K-1}p_1 & -p_{K-1}p_2 & \dots & p_{K-1}(1-p_{K-1}) \end{pmatrix} + \lambda \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{K-1} \end{pmatrix}$$

$$= \begin{pmatrix} p_1+\lambda w_1 & 0 & \dots & 0 \\ 0 & p_2+\lambda w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_{K-1}+\lambda w_{K-1} \end{pmatrix} - PP^T$$

$$\triangleq A - PP^T. \tag{37}$$

where $P = (p_1, p_2, \dots, p_{k-1})^T$. For notation convenience, we refer to p_k as $P(Y = k|X)$ for $k \in \{1, \dots, K\}$.

Now let us consider p_i, w_i as positive operators. G_λ and A are operator matrices, and P is an operator vector. From $G_\lambda = A - PP^T$, it is straightforward to calculate the inverse of G_λ by

$$G_\lambda^{-1} = A^{-1} + A^{-1}P(\mathbf{1} - P^TA^{-1}P)^{-1}P^TA^{-1}, \tag{38}$$

where $\mathbf{1}$ is an identity operator.

From (34), $\tilde{\boldsymbol{f}} - \boldsymbol{f}^0 = G_\lambda^{-1}(\boldsymbol{f}^0)Z_\lambda(\boldsymbol{f}^0)$. We write G_λ^{-1} as $G_\lambda^{-1} = RA^{-1}$, where $R = I + A^{-1}P(\mathbf{1} - P^TA^{-1}P)^{-1}P^T$. Clearly, for any $z \in \mathcal{H}$, we can have $\|G_\lambda^{-1}(\boldsymbol{f}^0)Z_\lambda(\boldsymbol{f}^0)\|_2^2 \leq \|G_\lambda^{-1}A\|_2^2\|A^{-1}Z_\lambda(\boldsymbol{f}^0)\|_2^2$. To find a bound for $\|A^{-1}z\|_2$, we introduce an operator matrix G_1 as $G_1 = \mathrm{diag}(p_1 - p_1^2 + \lambda w_1, \dots, p_{K-1} - p_{K-1}^2 + \lambda w_{K-1})$. Then $A^{-1}z$ can be written as $A^{-1}G_1G_1^{-1}z$. Regarding $A^{-1}G_1$, we have

Lemma 1. *Suppose $\forall i$, p_i, w_i, and $p_i - p_i^2$ are positive operators. Assume that*

$\|p_i^{-1}\| \sim K$ *and* $\|p_i\| \sim 1/K$. *For* $\lambda > 0$, *then* $\forall i$, $i \in \{1, 2, \ldots, K-1\}$,

$$\|(p_i + \lambda w_i - p_i^2)^{-1}(p_i + \lambda w_i)\|_2 \leq \frac{1}{1 - \|p_i\|_2} \ , \tag{39}$$

$$\|(p_i + \lambda w_i)^{-1}(p_i + \lambda w_i - p_i^2)\|_2 \leq 1 + \|p_i\|_2 \ . \tag{40}$$

By the definition of $Z_\lambda(\boldsymbol{f})$ and (25), it is known that $Z_\lambda(\boldsymbol{f}^0) = \lambda W \boldsymbol{f}^0$, Regarding $G_1^{-1} Z_\lambda(\boldsymbol{f}^0)$, we also have the following proposition.

Proposition 1. *Suppose* $p_i - p_i^2$ *and* w_i *are positive definite operators. If* $\forall i$, $\|p_i - p_i^2\|_2 \sim 1/K$, *then*

$$\|(p_i - p_i^2 + \lambda w_i)^{-1} \lambda w_i f_i^0\|_2^2 \leq M_i \lambda K \quad \textit{as } \lambda \to 0,$$

where M_i *is a constant independent of* λ *and* K.

4.3.1.2 Stochastic Error

To begin the analysis of the stochastic error, let us define $G_{n\lambda}(\boldsymbol{f}) \triangleq D^2 l_{n\lambda}(\boldsymbol{f})$. Ignoring the error in the linearzation, based on (26) and (27), we have

$$\begin{aligned} & Dl_{n\lambda}(\hat{\boldsymbol{f}}) - Dl_{n\lambda}(\tilde{\boldsymbol{f}}) = G_{n\lambda}(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) \\ & \Rightarrow \hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}} = -G_{n\lambda}^{-1}(\tilde{\boldsymbol{f}}) Dl_{n\lambda}(\tilde{\boldsymbol{f}}). \end{aligned} \tag{41}$$

The $G_{n\lambda}^{-1}(\tilde{\boldsymbol{f}})$ in (41) can be very complicated. It is not easy to calculate its close form. To overcome this difficulty, we introduce an intermediate function $\bar{\boldsymbol{f}}$, such that

$$\bar{\boldsymbol{f}} \triangleq \tilde{\boldsymbol{f}} - G_\lambda^{-1}(\tilde{\boldsymbol{f}}) Dl_{n\lambda}(\tilde{\boldsymbol{f}}). \tag{42}$$

Obviously, $\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}} = (\hat{\boldsymbol{f}} - \bar{\boldsymbol{f}}) + (\bar{\boldsymbol{f}} - \tilde{\boldsymbol{f}})$. Moreover, it can be shown that $\hat{\boldsymbol{f}} - \bar{\boldsymbol{f}}$ can be bounded in terms of $\bar{\boldsymbol{f}} - \tilde{\boldsymbol{f}}$.

Proposition 2.

$$\hat{\boldsymbol{f}} - \bar{\boldsymbol{f}} = -G_\lambda^{-1}(\tilde{\boldsymbol{f}}) \left[D^2 l_n(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{f}) - D^2 l(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) \right]. \tag{43}$$

Proof: From (41) and (42), we have

$$Dl_{n\lambda}(\tilde{\boldsymbol{f}}) + D^2 l_{n\lambda}(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) = 0,$$
$$Dl_{n\lambda}(\tilde{\boldsymbol{f}}) + D^2 l_{\lambda}(\tilde{\boldsymbol{f}})(\bar{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) = 0.$$

Using the above two equations, we can get

$$D^2 l_{n\lambda}(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) = D^2 l_{\lambda}(\tilde{\boldsymbol{f}})(\bar{\boldsymbol{f}} - \tilde{\boldsymbol{f}}).$$

Then

$$\begin{aligned} G_\lambda(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \bar{\boldsymbol{f}}) &= D^2 l_\lambda(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}} - (\bar{\boldsymbol{f}} - \tilde{\boldsymbol{f}})) \\ &= D^2 l_\lambda(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) - D^2 l_\lambda(\tilde{\boldsymbol{f}})(\bar{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) \\ &= D^2 l_\lambda(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) - D^2 l_{n\lambda}(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) \\ &= D^2 l(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) - D^2 l_n(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}). \end{aligned}$$

Therefore,

$$\hat{\boldsymbol{f}} - \bar{\boldsymbol{f}} = -G_\lambda^{-1}(\tilde{\boldsymbol{f}}) \left[D^2 l_n(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) - D^2 l(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) \right]. \qquad \square$$

Proposition 3. *Suppose the assumptions in Lemma 4 hold, then*

$$\|\hat{\boldsymbol{f}} - \bar{\boldsymbol{f}}\|_2 \leq o_K(1)\|\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}\|_2, \ \textit{as } n \to +\infty.$$

Proof: From Proposition 1, we know that,

$$\begin{aligned} \|\hat{\boldsymbol{f}} - \bar{\boldsymbol{f}}\|_2 &= \|G_\lambda^{-1}(\tilde{\boldsymbol{f}}) \left[D^2 l_n(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) - D^2 l(\tilde{\boldsymbol{f}})(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) \right] \|_2 \\ &\leq \|R\|_2 \|A^{-1} G_1\|_2 \|G_1^{-1}(D^2 l_n(\tilde{\boldsymbol{f}}) - D^2 l(\tilde{\boldsymbol{f}}))(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}})\|_2 \\ &\leq \|R\|_2 \|G_1^{-1}(D^2 l_n(\tilde{\boldsymbol{f}}) - D^2 l(\tilde{\boldsymbol{f}}))(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}})\|_2 \\ &\leq \|R\|_2 \|G_1^{-1}\|_2 \|D^2 l_n(\tilde{\boldsymbol{f}}) - D^2 l(\tilde{\boldsymbol{f}})\|_2 \|\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}\|_2, \end{aligned}$$

where $R = I + A^{-1}P(\mathbf{1} - P^T A^{-1} P)^{-1} P^T$.

First, we get the close form of $Dl_n(\boldsymbol{f})$ and $D^2l_n(\boldsymbol{f})$. By calculation, it is easy to show that

$$Dl_n(\boldsymbol{f}) = -2 \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}(y_{i1} - p_1(\boldsymbol{f}(x_i))) \\ \vdots \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}(y_{iK-1} - p_{K-1}(\boldsymbol{f}(x_i))) \end{bmatrix}. \tag{44}$$

Here $\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik} = n$ and $\forall i$, $\sum_{k=1}^{K} p_k(\boldsymbol{f}(x_i)) = 1$. Taking the Frechet derivative on $Dl_n(\boldsymbol{f})$, we obtain the close form of matrix $D^2l_n(\boldsymbol{f})$ as

$$[D^2l_n(\boldsymbol{f})]_{lk} = 2 \begin{cases} \frac{1}{n}\sum_{i=1}^{n} p_k(\boldsymbol{f}(x_i))[1 - p_k(\boldsymbol{f}(x_i))] & \text{if} \quad l = k \\ \\ -\frac{1}{n}\sum_{i=1}^{n} p_l(\boldsymbol{f}(x_i))p_k(\boldsymbol{f}(x_i)) & \text{if} \quad l \neq k \end{cases} \tag{45}$$

Denote $\tilde{g}_{kk}(x) = p_k(\boldsymbol{f}(x_i))[1 - p_k(\boldsymbol{f}(x_i))]$ and $\tilde{g}_{lk}(x) = -p_l(\boldsymbol{f}(x_i))p_k(\boldsymbol{f}(x_i))$ for $l \neq k$. Ignoring the constant 2 in (45), then

$$[D^2l_n(\tilde{\boldsymbol{f}}) - D^2l(\tilde{\boldsymbol{f}})]_{lk} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} \tilde{g}_{kk}(x_i) - E(\tilde{g}_{kk}(x)) & \text{if} \quad l = k \\ \\ \frac{1}{n}\sum_{i=1}^{n} \tilde{g}_{lk}(x_i) - E(\tilde{g}_{lk}(x)) & \text{if} \quad l \neq k \end{cases} \tag{46}$$

Obviously, $\forall x$, $0 \leq p_k(\boldsymbol{f}(x)) \leq 1$ and $0 \leq p_l(\boldsymbol{f}(x)) + p_k(\boldsymbol{f}(x)) \leq 1$ when $l \neq k$. Hence, for any l and k, $0 \leq \tilde{g}_{lk}(x) \leq 1/4$. A direct computation of expectation shows that

$$E(\|[D^2l_n(\tilde{\boldsymbol{f}}) - D^2l(\tilde{\boldsymbol{f}})]_{lk}\mu\phi\|_2^2) \leq O(n^{-1})\|\mu\|_2^2\|\phi\|_2^2. \tag{47}$$

Clearly, L_2 norm can be bounded by the Frobenuis norm (i.e., if A is a $m \times n$ matrix, then $||A||_2 \leq ||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$.). Using this fact and (47), we can get

$$E(\|D^2l_n(\tilde{\boldsymbol{f}}) - D^2l(\tilde{\boldsymbol{f}})\|_2^2) \leq O(n^{-1}). \tag{48}$$

Recall that $G_1(f) = \text{diag}(p_1(\boldsymbol{f})(1 - p_1(\boldsymbol{f})) + \lambda w_1, \ldots, (p_{K-1}(\boldsymbol{f})(1 - p_{K-1}(\boldsymbol{f})) + \lambda w_{K-1}) \triangleq \text{diag}(u_1 + \lambda w_1, \ldots, u_{K-1} + \lambda w_{K-1})$, where $u_k = p_k(\boldsymbol{f})(1 - p_k(\boldsymbol{f})$. Based on Lemma 5 and Lemma 8, $\forall i$, $||(u_i + \lambda w_i)^{-1}||_b^2 = O_p(K^2(\lambda K)^{-(b+1/2m)})$. Note that

$b = 0$ for L_2 norm, then clearly $\|G_1^{-1}\|_2^2 \leq O(K^2\lambda^{-1/(2m)})$. From Lemma 4, it has shown that $\|R\|_2 = O(K)$. Combining these statements, we obtain,

$$\begin{aligned}\|\hat{\boldsymbol{f}} - \bar{\boldsymbol{f}}\|_2 &\leq \|R\|_2\|G_1^{-1}\|_2\|D^2 l_n(\tilde{\boldsymbol{f}}) - D^2 l(\tilde{\boldsymbol{f}})\|_2\|\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}\|_2 \\ &\leq O(K^{2-1/(4m)})O(\lambda^{-1/(4m)})O(n^{-1/2})\|\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}\|_2. \end{aligned} \tag{49}$$

Hence, we conclude that $||\hat{\boldsymbol{f}} - \bar{\boldsymbol{f}}||_2 \leq o_K(1)||\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}||_2$ as $n \to \infty$. □

From (26), $Dl_{n\lambda}(\tilde{\boldsymbol{f}}) = Dl_{n\lambda}(\tilde{\boldsymbol{f}}) - Dl_{\lambda}(\tilde{\boldsymbol{f}})$, and using (42),

$$\bar{\boldsymbol{f}} - \tilde{\boldsymbol{f}} = -G_{\lambda}^{-1}(\tilde{\boldsymbol{f}})[Dl_n(\tilde{\boldsymbol{f}}) - Dl(\tilde{\boldsymbol{f}})]. \tag{50}$$

As mentioned, $\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}$ can be decomposed as $\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}} = (\hat{\boldsymbol{f}} - \bar{\boldsymbol{f}}) + (\bar{\boldsymbol{f}} - \tilde{\boldsymbol{f}})$. Based on Proposition 2 and 3, the property of $\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}$ will be further studied to derive the upper bound for $\hat{g}_r - \tilde{g}_r$ in Section 4.3.3.

4.3.2 Analysis of $\tilde{M}$

With a little abuse of notation, hereafter we do not clearly distinguish $\vec{g}_r$ and g_r, $\vec{\hat{g}}_r$ and $\hat{g}_r$, or $\vec{\tilde{g}}_r$ and $\tilde{g}_r$. In this section, we focus to obtain an upper bound of $\frac{1}{\sqrt{n}}\|\tilde{g}_r - g_r^0\|_2$.

Let us start from $\tilde{\boldsymbol{f}}$ and $\boldsymbol{f}^0$. Note that we have built a relation between function vectors $\tilde{\boldsymbol{f}}$ and $\boldsymbol{f}^0$ in (34), i.e.,

$$\tilde{\boldsymbol{f}} - \boldsymbol{f}^0 = -G_{\lambda}^{-1}(\boldsymbol{f}^0)(\lambda W \boldsymbol{f}^0) \triangleq A_1(\lambda)\boldsymbol{f}^0, \tag{51}$$

where $A_1(\lambda)$ is defined as $A_1(\lambda) = -G_{\lambda}^{-1}\lambda W$. According to the definition of $\tilde{M}$ and M_0, we have

$$\begin{aligned}\tilde{M} = \frac{1}{n}\tilde{F}\tilde{F}^T &= M^0 + \frac{1}{n}F^0 H (F^0)^T \\ &= M^0 + M^*, \end{aligned} \tag{52}$$

where $H = A_1^T(\lambda) + A_1(\lambda) + A_1^T(\lambda)A_1(\lambda)$, and $M^* = \frac{1}{n}F^0 H (F^0)^T$. To investigate the convergence property of $\frac{1}{\sqrt{n}}\|\tilde{g}_r - g_r^0\|_2$, we concentrate on the relation of eigenvectors of $\tilde{M}$ and M^0, where these two matrices are not very different in some sense. The

common techniques used here are derived from the perturbation theory (Kato, 1966). Basically, it considers the changes of eigenvectors when passing over from M^0 to $M^0 + M^*(M^* = \tilde{M} - M^0)$. Some useful lemmas (Kneip, 1994) are listed in Appendix B without giving proofs. We also impose some regularity conditions on $\sum_{j=1}^{K-1} \beta_{jr}^2$, the eigenvalue of matrix M^0.

Assumption 1. *There exists a constant D_1 such that, for all $r, s \in \{1, , \ldots, L_0\}$, $r \neq s$,*

$$\left| \sum_{j=1}^{K-1} \beta_{jr}^2 - \sum_{j=1}^{K-1} \beta_{js}^2 \right| \geq D_1 \sum_{j=1}^{K-1} \beta_{jr}^2.$$

This assumption is just a technical requirement. we assume that the eigenvalue $\sum_{j=1}^{K-1} \beta_{jr}^2$ decreases rapidly as r increases.

Suppose an eigenvector of M^0 is g_r^0. The corresponding eigenvalue of the eigenvector g_r^0 is l_r. From Lemma 9 and Lemma 10, we know that there exists a real eigenvector $\tilde{g}_r$ of $\tilde{M}$, such that,

$$\frac{1}{\sqrt{n}} \|\tilde{g}_r - g_r^0\| \leq \alpha(l_r)_2 \frac{2}{1 - 4\beta}, \tag{53}$$

where $\alpha(l_r)_2 \leq \|S(l_r)\| \cdot \|M^* U(l_r)\|$. Here $U(l_r)$ is the projection matrix projecting onto the eigenspace of M^0 for l_r, i.e., $U(l_r) = \frac{1}{n} g_r^0 (g_r^0)^T$, and $\boldsymbol{S}(l_r)$ is the reduced resolvent of M^0 for l_r, i.e.,

$$\boldsymbol{S}(l_r) = \sum_{\tau \neq l_r, \tau \in \mathcal{EG}(M^0)} \frac{1}{\tau - l_r} U(l_r),$$

where $\mathcal{EG}(M^0)$ is the set of all eigenvalues of M^0. First we compute the bound for $\alpha(l_r)_2$, which involves $M^* U(l_r)$ and $S(l_r)$. In particular,

$$\begin{aligned} \|M^* U(l_r)\| = \frac{1}{n} \|F^0 H (F^0)^T \frac{1}{\sqrt{n}} g_r^0 \frac{1}{\sqrt{n}} (g_r^0)^T\| &= \frac{1}{n} \|F^0 H (F^0)^T \frac{1}{\sqrt{n}} g_r^0\| \\ &\leq \frac{1}{n} \|H\| \|F^0\| \|(F^0)^T\|. \end{aligned}$$

Furthermore,

$$\|M^*U(l_r)\|_2^2 \leq \|H\|_2^2 \left[\lambda_{max}(\frac{1}{n}F^0(F^0)^T)\right]^2, \tag{54}$$

where $\lambda_{max}(\cdot)$ is the notation for the largest eigenvalue.

From Lemma 12, we get $\|H\|_2^2 \sim \lambda K$. The proof is given in the appendix. Based on Assumption 1, it implies that $\|S(l_r)\|_2 \leq 1/(D_1 l_r)$. Thus,

$$\alpha(l_r)_2^2 \leq \frac{(\lambda K)[\lambda_{max}(\frac{1}{n}F^0(F^0)^T)]^2}{(D_1 l_r)^2} \sim O(\lambda K). \tag{55}$$

Note that $M^0 = \frac{1}{n}F^0(F^0)^T$. According to the construction mechanism of M, we can see that the eigenvalues of M^0 is with order K, i.e., $\|\lambda(M^0)\| \sim O(K)$. Next we need to verify the conditions for $\beta(l_r)$, which is defined in Lemma 10. As the tuning parameter λ goes to zero, it is easy to satisfy the condition that $\|S(l_r)\|_2 \cdot \|M^*U(l_r)\|_2 \leq (\lambda K)\lambda_{max}(M^0) \leq \frac{1}{8}$. Considering the bound of $\|M^*S(l_r)\|$, based on its definition, we can get

$$\|M^*S(l_r)\|_2 \leq \sum_{\tau \neq l_r} \frac{1}{|\tau - l_r|}\|M^*U(l_r)\|, \tag{56}$$

Hence, $\|M^*S(l_r)\|_2 \leq \frac{1}{8}$ as $\lambda \to 0$. Now we have

$$\beta_{l_r} = \max\{\|M^*S(l_r)\|, \|S(l_r)\| \cdot \|M^*U(l_r)\|\} \leq \frac{1}{8}. \tag{57}$$

Applying the results (55) and (57) into (53) through Lemma 10, we obtain the following theorem.

Theorem 2. *Suppose the assumption 1 holds in addition to the conditions in Theorem 1, then we obtain that*

$$\frac{1}{\sqrt{n}}\|\tilde{g}_r - g_r^0\|_2^2 \leq O(\lambda K). \tag{58}$$

4.3.3 Analysis of $\hat{M}$

Recall that $\hat{M} = \frac{1}{n}\hat{F}\hat{F}^T$, where $\hat{F}$ is defined in the same fashion as F in (30). For notation convenience, we denote the vector function $\hat{\boldsymbol{f}}(x) = (f_1(x), \ldots, f_{K-1}(x))^T$. To understand the relation between the matrices $\hat{M}$ and $\tilde{M}$, we come to study the relation between $\hat{\boldsymbol{f}}$ and $\tilde{\boldsymbol{f}}$. By the fact that $(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) - (\hat{\boldsymbol{f}} - \bar{\boldsymbol{f}}) = (\bar{\boldsymbol{f}} - \tilde{\boldsymbol{f}})$ and using (42) and (43), we can get that

$$\left(I + G_\lambda^{-1}(\tilde{\boldsymbol{f}})[D^2 l_n(\tilde{\boldsymbol{f}}) - D^2 l(\tilde{\boldsymbol{f}})]\right)(\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}}) = (\bar{\boldsymbol{f}} - \tilde{\boldsymbol{f}}). \tag{59}$$

Combining (50) and (59), we can obtain that

$$\begin{aligned}\hat{\boldsymbol{f}} - \tilde{\boldsymbol{f}} &= -\left(I + G_\lambda^{-1}(\tilde{\boldsymbol{f}})[D^2 l_n(\tilde{\boldsymbol{f}}) - D^2 l(\tilde{\boldsymbol{f}})]\right)^{-1} G_\lambda^{-1}(\tilde{\boldsymbol{f}})[Dl_n(\tilde{\boldsymbol{f}}) - Dl(\tilde{\boldsymbol{f}})] \\ &\triangleq B \cdot d\tilde{\boldsymbol{f}},\end{aligned} \tag{60}$$

where $B = -\left(I + G_\lambda^{-1}(\tilde{\boldsymbol{f}})[D^2 l_n(\tilde{\boldsymbol{f}}) - D^2 l(\tilde{\boldsymbol{f}})]\right)^{-1} G_\lambda^{-1}(\tilde{\boldsymbol{f}})$ and $d\tilde{\boldsymbol{f}} = Dl_n(\tilde{\boldsymbol{f}}) - Dl(\tilde{\boldsymbol{f}})$. Then we can build the connection between $\hat{F}$ and $\tilde{F}$, i.e.,

$$\begin{aligned}\hat{F} &= \begin{pmatrix} \hat{\boldsymbol{f}}^T(x_1) \\ \vdots \\ \hat{\boldsymbol{f}}^T(x_n) \end{pmatrix} = \begin{pmatrix} (\tilde{\boldsymbol{f}}(x_1) + B \cdot d\tilde{\boldsymbol{f}}(x_1))^T \\ \vdots \\ (\tilde{\boldsymbol{f}}(x_n) + B \cdot d\tilde{\boldsymbol{f}}(x_n))^T \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{f}}^T(x_1) \\ \vdots \\ \tilde{\boldsymbol{f}}^T(x_n) \end{pmatrix} + (B \cdot D\tilde{F})^T \\ &= \tilde{F} + (B \cdot D\tilde{F})^T,\end{aligned} \tag{61}$$

where $D\tilde{F} = (d\tilde{\boldsymbol{f}}(x_1), \ldots, d\tilde{\boldsymbol{f}}(x_n))^T$. From the definition of $\hat{M}$, we have

$$\begin{aligned}\hat{M} &= \frac{1}{n}\hat{F}\hat{F}^T = \frac{1}{n}[\tilde{F} + (B \cdot D\tilde{F})^T][\tilde{F} + (B \cdot D\tilde{F})^T]^T \\ &= \frac{1}{n}\tilde{F}\tilde{F}^T + \frac{1}{n}[\tilde{F}(B \cdot D\tilde{F}) + (B \cdot D\tilde{F})^T\tilde{F}^T + (B \cdot D\tilde{F})^T(B \cdot D\tilde{F})] \\ &= \tilde{M} + \frac{1}{n}[\tilde{F}(B \cdot D\tilde{F}) + (B \cdot D\tilde{F})^T\tilde{F}^T + (B \cdot D\tilde{F})^T(B \cdot D\tilde{F})].\end{aligned} \tag{62}$$

Using (62), we come to study the relation on the eigenvectors of $\hat{M}$ and $\tilde{M}$. Based on the definition of $\tilde{M}$, we can see that $\tilde{M}$ is symmetric and has at most L nonzero eigenvalues. Suppose there is a singular value decomposition (SVD) of $\tilde{M}$ such that

$Q^T\tilde{M}Q = \tilde{\Lambda}$, where $\tilde{\Lambda} = \text{diag}(\eta_1(\tilde{\Lambda}), \ldots, \eta_L(\tilde{\Lambda}))$ is a diagonal matrix with $\eta_1(\tilde{\Lambda}) \geq \ldots, \geq \eta_L(\tilde{\Lambda})$. Here Q is a $n \times L$ matrix with $Q^TQ = I$. It is easy to see that the nonzero eigenvalues of $\tilde{M}$ and of $\tilde{\Lambda}$ are identical. Furthermore, if $\gamma_r(\tilde{\Lambda})$ is the eigenvector of $\tilde{\Lambda}$, then

$$\frac{1}{\sqrt{n}}\tilde{g}_r = Q\gamma_r(\tilde{\Lambda}).$$

The derivation is that if there exists a vector y such that $\tilde{\Lambda}y = \eta y$, then we can have

$$\begin{aligned}\tilde{M}Qy &= Q\tilde{\Lambda}Q^TQy = Q\tilde{\Lambda}y \\ &= Q\eta y = \eta(Qy).\end{aligned}$$

It means that Qy is the eigenvector of $\tilde{M}$. By defining $\hat{\Lambda} = Q^T\hat{M}Q$, then consequently,

$$\begin{aligned}\frac{1}{n}\|\hat{g}_r - \tilde{g}_r\|_2^2 &= \|Q\gamma_r(\hat{\Lambda}) - Q\gamma_r(\tilde{\Lambda})\|_2^2 \\ &= \|\gamma_r(\hat{\Lambda}) - \gamma_r(\tilde{\Lambda})\|_2^2. \end{aligned} \tag{63}$$

The last equality is provided by the orthogonality of Q. Now we can write

$$\begin{aligned}\hat{\Lambda} - \tilde{\Lambda} &= Q^T(\hat{M} - \tilde{M})Q \\ &= \frac{1}{n}Q^T[\tilde{F}(B \cdot D\tilde{F}) + (B \cdot D\tilde{F})^T\tilde{F}^T + (B \cdot D\tilde{F})^T(B \cdot D\tilde{F})]Q \\ &\triangleq \Xi.\end{aligned}$$

Let us denote ξ_{ij} as the ijth element of matrix Ξ. From Lemma 14, we have $E\xi_{rs}^2 = O(K\frac{K}{n}(\lambda K)^{-\frac{1}{2m}})$. The proof of the lemma is given in Appendix B. Obviously, we can find a $\delta_K = \frac{1}{n}(\lambda K)^{-\frac{1}{2m}}$ satisfying properties $K\delta_K \to 0$ as $n \to \infty$ and $\sup_{r,s}(E\xi_{rs})^2/(\tilde{\lambda}_r\tilde{\lambda}_s)) = O(\delta_K)$ as $n \to \infty$.

Suppose an eigenvector of $\tilde{M}$ is $\tilde{g}_r$. The corresponding eigenvalue of the eigenvector $\tilde{g}_r$ is $\tilde{\lambda}_r$. Based on Lemma 13, we can know that there exists an eigenvector $\hat{g}_r$ of $\hat{M}$, such that,

$$\frac{1}{n}\|\hat{g}_r - \tilde{g}_r\|_2^2 = O\left(\frac{\sum_{s\neq r}\left[E\xi_{rs}^2/\max\{\tilde{\lambda}_r, \tilde{\lambda}_s\}\right]}{\tilde{\lambda}_r}\right),\ r = 1, \ldots, L, \tag{64}$$

where $\tilde{\lambda}_r, \tilde{\lambda}_s$ are eigenvalues of $\tilde{M}$. Recall the construction mechanism of $\tilde{M}$, we know $\tilde{M}$'s eigenvalue $\lambda_r(\tilde{M})$ has order K, i.e., $\lambda_r(\tilde{M}) = \lambda_r(\tilde{\Lambda}) = O(K)$. Therefore, we get

$$O\left(\frac{\sum_{s\neq r}\left[E\xi_{rs}^2/\max\{\tilde{\lambda}_r,\tilde{\lambda}_s\}\right]}{\tilde{\lambda}_r}\right) = O(\frac{\sum_{s\neq r} K\frac{K}{n}(\lambda K)^{-\frac{1}{2m}}/K}{K}) = \frac{L}{n}(\lambda K)^{-\frac{1}{2m}}. \quad (65)$$

Hence we get the following theorem.

Theorem 3. *Suppose the assumptions in Lemma 13 and Lemma 14 hold, then we can obtain that*

$$\frac{1}{n}\|\hat{g}_r - \tilde{g}_r\|_2^2 = \frac{L}{n}(\lambda K)^{-\frac{1}{2m}}. \quad (66)$$

Theorem 1 follows immediately from Theorem 2 and Theorem 3. By equating the asymptotic orders of $\frac{1}{n}\|\tilde{g} - g^0\|_2^2$ and $\frac{1}{n}\|\hat{g} - \tilde{g}\|_2^2$, we can choose $\lambda_n = \frac{1}{K}(n/L)^{-\frac{2m}{2m+1}}$. Using the fact $\|\hat{g} - g^0\| \leq (\|\hat{g} - \tilde{g}\|) + \|\tilde{g} - g^0\|$, then

$$\frac{1}{n}\|\hat{g}_r - g_r^0\|_2^2 \leq \left(\frac{n}{L}\right)^{-\frac{2m}{2m+1}}. \quad (67)$$

According to the factor model for $\hat{\boldsymbol{f}}$ in (29), we get an upper bound of convergence rate of the penalized likelihood estimator as

$$\frac{\sum_{k=1}^K \|\hat{f}_k - f_k^0\|_2^2}{K} \leq C\left(\frac{n}{L}\right)^{-\frac{2m}{2m+1}},$$

where C is a constant independent on n, K and L.

4.4 Simulation Example

To illustrate the performance of the proposed factor model on classifier functions, we consider a simulation example with 2 factors and $K = 10$ classes. The two factor functions on $[0, 1]$ are

$$g_1(x) = (2x-1)^2, \quad g_2(x) = \frac{\sin(2\pi x)}{2 - \sin(2\pi x)}.$$

The classifier functions are

$$f_k(x) = \frac{k}{10}g_1(x) + \frac{10-k}{10}g_2(x),\ k = 1, \ldots, 9,$$

and with a baseline function $f_{10}(x) = 1$. To simulate the data, we randomly generate $N = 500$ sample of x from $[0, 1]$, and get their response $Y \in \{1, \ldots, K\}$ based on the multi-logit model in (19).

To measure the accuracy of the estimated classifier functions, we use the integrated squared error $ISE = E_X\{\hat{\boldsymbol{f}}(x) - \boldsymbol{f}(x)\}^2$. For each replication of the simulation, the ISE is calculated by Monte Carlo integration using 1000 test points randomly drawn from $[0, 1]$. we compare the proposed factor model with multi-logit model for 100 simulation study. The resulting average integrated squared errors and its corresponding standard errors (in parentheses) are shown in Table 3. The estimated classifier functions from one simulation run is shown in Figure 4.4. We can see that the classifier function estimated from the factor model is more accurate than those estimated from the multi-logit model.

Table 3: Comparison of estimated classifier functions.

	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$
factor model	0.192 (0.014)	0.169 (0.010)	0.183 (0.013)	0.163 (0.016)	0.170 (0.012)
multi-logit	0.223 (0.014)	0.194 (0.012)	0.213 (0.015)	0.209 (0.019)	0.221 (0.015)
	$f_6(x)$	$f_7(x)$	$f_8(x)$	$f_9(x)$	
factor model	0.158 (0.012)	0.163 (0.014)	0.161 (0.012)	0.156 (0.012)	
multi-logit	0.188 (0.012)	0.203 (0.014)	0.201 (0.013)	0.187 (0.012)	

4.5 Discussion

In this work, we proposed a factor model for classifier functions under the multi-logit models. We have shown an upper bound of the convergence rate of classifier function from the factor model. It is seen that the upper bound mainly depends on the number of factors, which implies that the proposed method can achieve better classification accuracy when there is a large number of categories.

We have not studied the choice of the number of factors for the proposed method. A common approach (Kneip, 1994, Li, 1991) is to use the hypothesis testing to estimate the dimensionality of the factor model. The sum of $(K-L)$ smallest eigenvalues

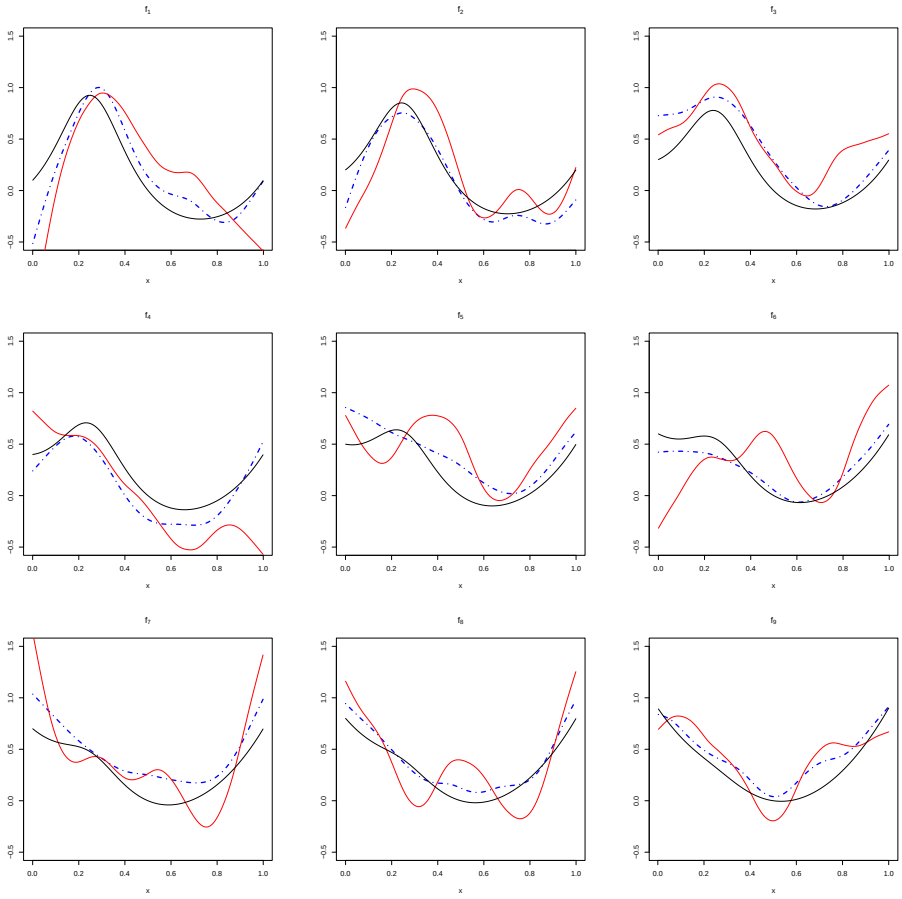


Figure 20: The comparison of the estimated classifier function in one run example (Black line: the true classifier function; Red line: the estimated classifier function from multi-logit; Blue dashed line: the estimated classifier function from factor model.).

of matrix $\hat{M}$ can be used to develop a related test statistic. Then the hypothesis can provide a guideline to choose the smallest L which the hypothesis is not rejected. It can be a topic for our future work.

We have focused on the asymptotic properties of the factor model on classifier functions for multi-logit models in this work. The concept of the factor model can also be applied to other multi-category classification methods such as multi-category support vector machines (Lee et al., 2002). It can be more challenging to study its asymptotic properties for factor models due to more complicated loss functions.

CHAPTER V

A STATISTICAL APPROACH TO QUANTIFYING THE ELASTIC DEFORMATION OF NANOMATERIALS

5.1 Introduction

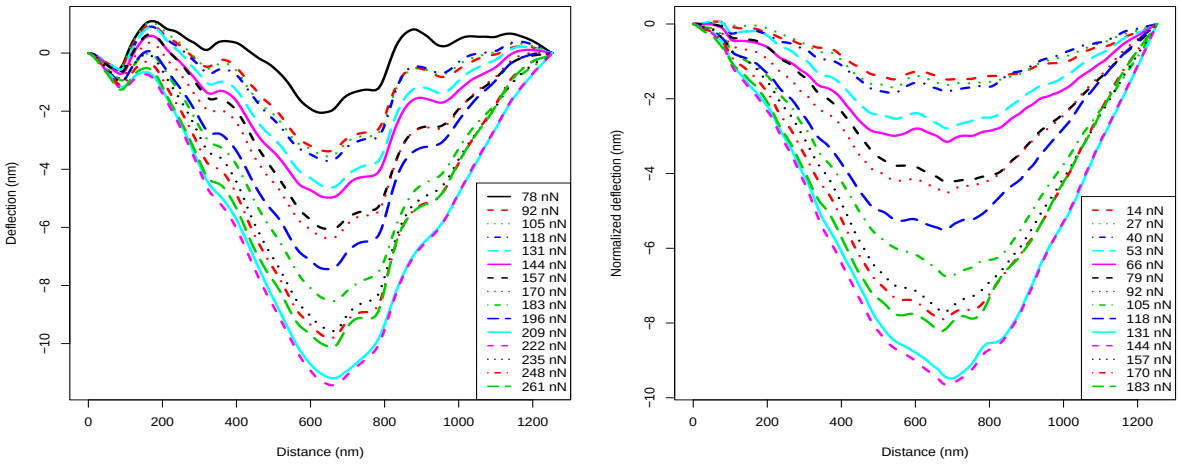
Nanotechnology has provided unprecedented understanding and applications on materials and is impacting many fields through the development of nanodevices and nanosystems that exhibit superior performances. The fundamental building blocks in constructing such devices and systems are one-dimensional (1D) nanomaterials, such as carbon nanotubes, semiconductor nanowires, and oxide nanobelts. The mechanical behavior of 1D nanomaterials is one of most important properties dictating their applications in nanotechnology. Among the several developed methods for measuring the elastic deformation properties of nanomaterials (Wong et al., 1997, Yu et al., 2000, and Poncharal et al., 1999), one approach to quantifying the elastic modulus of 1D nanomaterials is based on the atomic force microscopy (AFM). A common strategy is to deform a 1D nanostructure using an AFM tip, which pushes the 1D nanostructure at some locations. Then the elastic modulus is determined through quantifying the force-displacement curve. The accuracy of this measurement is, however, limited by noise factors such as the size of the tip, the accuracy of positioning the AFM tip on the object, the surface roughness of the 1D nanomaterials, and the stability of the structure during measurements. New approaches are needed for analyzing the data received from nano-scale measurements, so that the derived information can be reliably used to characterize the mechanical properties of nanomaterials. The objective of this article is to propose a new approach for quantitative nanomechanics through statistical and physical modeling.

Recently, Mai and Wang (2006) proposed a new approach for quantifying the elastic deformation behavior of 1D nanostructures. The approach is based on a continuous deformation of a Znic Oxide (ZnO) nanobelt, which is supported at its two ends by a trenched substrate, using an AFM tip in contact mode. The AFM tip scans along the length of the nanobelt under a constant applied force, and thus the segment across the trench is deformed. A quantitative fitting of the force-deflection curve is used for estimating the elastic modulus of the nanobelt. However, the measured data are largely affected by the imperfect shape of the nanobelt, its surface roughness, size and shape of the AFM tip, and the instability of the measurement technique at such a small scale. Moreover, the level of allowable tolerance on measurement errors for the nanomaterials decreases since noise or error becomes much larger in comparison to the small response signals from the nanomaterials. The data analysis is complicated by a lack of confidence in the assumed physical model to accommodate the uncertainty in the contact between the nanobelt and the supporting trench. One possible physical model is the simply-supported beam model (SSBM) (Benham and Crawford, 1987). The SSBM is an ideal case that does not account for the various experimental uncertainties and artifacts. In this paper, we use an empirical statistical modeling technique to identify the effects of these artifacts and their influence on data analysis. After filtering out such effects, we can accurately, reliably and efficiently determine the elastic modulus based on the physical law. Our study sets a good and early example for quantitative nanomechanics. The proposed methodology can be extended to other fields in nanotechnology such as nanoelectronics and nanomeasurements.

5.2 Existing Method

Mai and Wang (2006) used a physical vapor deposition method to synthesize the ZnO nanobelts with a rectangular cross-section. A silicon substrate was prepared with long and parallel trenches carved at its surface by nanofabrication. The trenches are about

200 nm deep and 1.25 m wide. They manipulated the long ZnO nanobelts across the trenches over many periods. A scanning electron microscopy (SEM) and AFM were used to capture the morphology and dimensions of the nanobelt. The length and width of the nanobelt are captured by the SEM image and the thickness of the nanobelt is obtained from AFM image. In the mechanical measurement, an AFM tip scanned the nanobelt along its length direction in contact mode at a constant applied force. By changing the magnitude of the contact force from low to high, they obtained a series of bending profiles of the nanobelt.



(a) (b)

Figure 21: (a) The AFM image profiles of the suspended NB under different load forces in contact mode. (b) The normalized AFM image profile by subtracting the profile acquired at 78 nN from the profiles in (a).

The profiles of a suspended nanobelt along the length direction under different contact forces are shown in Figure 21(a). As shown in Figure 21(a), the image profiles of a nanobelt (denoted by NB) recorded the deflection of all the points along its length under different applied forces. Each curve was obtained by averaging ten consecutive measurements along its length under the same loading force. The curves in Figure 21(a) are not smooth due to a small surface roughness (around 1 nm) of the nanobelt. In addition, the as-attached nanobelt on the trenches is not perfectly straight, possibly

due to initial bending during the sample manipulation. Figure 21(a) indicates that there are some noise factors affecting the deflection curves. In order to eliminate the effect of the surface roughness and initial bending of the nanobelt (collectively referred to as initial bias), Mai and Wang (2006) proposed to calibrate the deflection curves by subtracting the initial profile (i.e., the profile measured under the lowest applied force of 78 nN) from those measured at higher applied forces. The normalized AFM image profiles are shown in Figure 21(b). A normalized force is obtained by subtracting 78 nN from the applied forces (see the inset box in Figure 21(b)).

Mai and Wang (2006) suggested the simply-supported beam model (SSBM) to quantify the elastic deflection (which they called the free-free beam model). The diagram of the SSBM is shown in Figure 22.

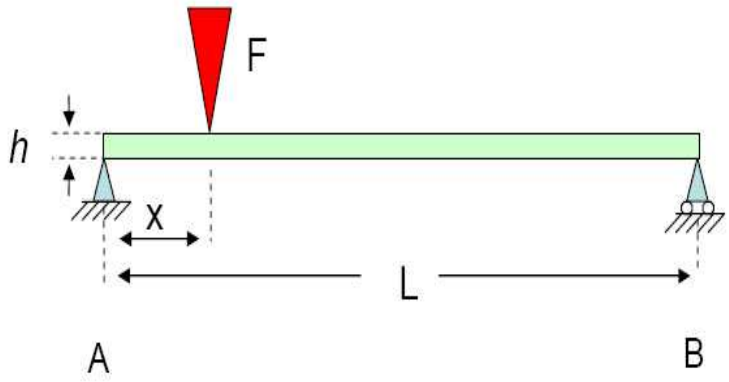


Figure 22: The schematic diagram of the simply-supported beam model (SSBM).

When a concentrated load force F is applied at the contact point x of the AFM tip away from the end A, the deflection of nanobelt at x is determined by

$$v = -\frac{Fx^2(L-x)^2}{3EIL}, \tag{68}$$

where E is the elastic modulus, L is the width of trench, and I is the moment of inertia given by $wh^3/12$ for the rectangular beam, where w and h are respectively the width and thickness of nanobelt. The notation in (68) is slightly different from that in

Mai and Wang (2006). Figure 23 shows an illustrative example of the SSBM profiles, which are symmetric and perfectly smooth but do not account for the noise factors in the measurements. The elastic modulus E is estimated by fitting the normalized AFM image profiles to the SSBM. Hereafter, we denote the method by Mai and Wang as the MW method.

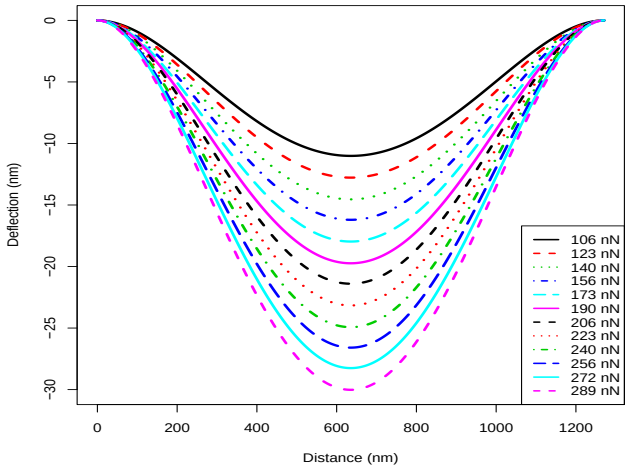


Figure 23: An example of SSBM profiles.

The elastic beam model for the bridged nanowire configuration is widely accepted in nanomechanics (see Salvetat et al., 1999 and Wu et al., 2005, etc). The current nanobelt experiment is in a linear elastic deflection region because the nanobelt has a maximum deflection change around 10 nm under the largest load force during the experiment, which is less than 1% deflection comparing to the length of the nanobelt (1250 nm). This approach has also been effectively used elsewhere such as Paulo et al. (2005). The simply-supported boundary assumption is validated in Mai and Wang (2006), which showed that SSBM fits the data better than the clamped-clamped beam model. It is also confirmed experimentally in their work.

5.2.1 Problem with the MW Method

By subtracting the profile acquired at 78 nN from the data, the shape of the normalized AFM image profiles in Figure 21(b) looks closer to the SSBM than that of the original profiles in Figure 21(a). This should give a better estimate of the elastic modulus E. On the other hand, if the initial profile behaves poorly, then subtracting this profile to normalize the data can result in poor estimation of E from the SSBM.

Recall that the deflection v in the SSBM in (68) is a linear function of the applied force F given the distance x. The reason for normalizing the data by subtracting the initial profile is to eliminate the initial bias. However, if some systematic errors due to imperfect boundary conditions and other unknown factors (collectively referred to as systematic bias) occurred during the experiment, normalizing the data may not be enough for obtaining a good fitting based on the SSBM. For example, in Figure 21(a) the deflection profiles under the applied force $F = 235, 248$ and 261 nN lie above those under the lower force $F = 209$ and 222 nN. This is inconsistent with the model equation in (68) because the deflection is expected to increase with force. The SSBM itself cannot explain this phenomenon. One possible explanation is the change of the boundary conditions, which can be nonlinear and irreversible during the measurement. This pattern still persists in the normalized profiles in Figure 21(b). Therefore, the MW method cannot be used to fit the profile data properly. It requires a more general model to identify other factors besides the initial bias.

To overcome these problems, we propose a physical-statistical model that integrates SSBM with a regression model. The regression model captures the initial bias and potential systematic biases introduced during measurement. We use model selection to identify terms associated with the systematic biases and adjust the profiles by subtracting these terms from the original profiles. This provides a better estimate of the elastic modulus E. We call the method *sequential profile adjustment by regression (SPAR).*

5.3 General Model and Model Selection

5.3.1 General Model

As shown in Figure 21(a), suppose there are K image profiles, i.e., the nanobelt is scanned sequentially under K different applied forces $F_1, F_2, \ldots, F_K$. The experimenter usually changes the magnitude of applied force F from low to high, i.e., $F_1 < F_2 < \cdots < F_K$. Each profile contains n points which are recorded at the distances of $x_1, x_2, \ldots, x_n$. We denote the deflection at the distance x under the applied force F as $v(x, F)$. Then the SSBM can be written as

$$v(x, F) = \beta(x)F, \tag{69}$$

where $\beta(x) = x^2(L-x)^2/(-3EIL)$. Let $\delta_0(x)$ be the *initial bias* and $\delta_k(x)$ for $k \geq 1$ be the *systematic bias* introduced when an AFM tip scans the NB along its length at the applied force F_k. The initial bias can be due to the surface roughness and initial bending. The systematic biases can be due to the uncertainty of boundary conditions, causing the occasional stick-slip events that occur at the ends of the nanobelt. The wear and tear of AFM tip and the nanobelt surface, the lateral shifting and sliding, and other artifacts can also be the causes. Such causes can occur at any stage of the experiment. These random causes cannot be effectively captured using deterministic mechanistic models, whereas they can be easily incorporated using statistical models. Thus we propose to model the deflection at scanned under the k-th applied force F_k as

$$v(x, F_k) = \beta(x)F_k + \delta_0(x) + \delta_1(x)I(k > 1) + \cdots + \delta_{K-1}(x)I(k > K-1) + \varepsilon(x, F_k), \tag{70}$$

where $I(\cdot)$ is an indicator function and $\varepsilon(x, F_k)$ is the error term. Note that the indicator function is to model the sequential nature of the experiment. Specifically, when the force F_k is applied to make the AFM tip in contact with the NB, the

proposed approach models the deflection as

$$v(x, F_k) = \beta(x)F_k + \delta_0(x) + \delta_1(x) + \cdots + \delta_{k-1}(x) + \varepsilon(x, F_k). \tag{71}$$

In reality, there may or may not be a bias at stage k, i.e., some of the δ_k's may be zero. We therefore use a model selection technique to identify the significant δ_k's and include only them in the final model.

5.3.2 Model Selection

The general model (71) considers all potential bias factors. In reality, it is likely that only a few of them contribute toward the deflection on the nanobelt. So it is important to find significant δ_k's and build an appropriate model. Given the distance xi, the model (71) is a linear regression with $K+1$ parameters $1/E, \delta_0(x_i), \delta_1(x_i), \ldots, \delta_{K-1}(x_i)$, i.e.,

$$\begin{pmatrix} v(x_i, F_1) \\ v(x_i, F_2) \\ \vdots \\ v(x_i, F_K) \end{pmatrix} = \begin{pmatrix} \gamma(x_i)F_1 & 1 & 0 & \ldots & 0 \\ \gamma(x_i)F_2 & 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma(x_i)F_K & 1 & 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} 1/E \\ \delta_0(x_i) \\ \vdots \\ \delta_{K-1}(x_i) \end{pmatrix} + \boldsymbol{\varepsilon}_i, \tag{72}$$

where $\gamma(x_i) = x_i^2(L - x_i)^2/(-3IL)$ incorporates the knowledge of the SSBM.

In the error vector $\boldsymbol{\varepsilon}_i = (\varepsilon(x_i, F_1), \ldots, \varepsilon(x_i, F_K))^T$, $\varepsilon(x_i, F_k)$ represents the error occurred at distance x_i under applied force F_k. The model (71) considering all x_i is an over-parameterized linear model with parameters $1/E, \boldsymbol{\delta}_0(\boldsymbol{x}), \boldsymbol{\delta}_1(\boldsymbol{x}), \ldots, \boldsymbol{\delta}_{K-1}(\boldsymbol{x})$ where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, $\boldsymbol{\delta}_k(\boldsymbol{x}) = (\delta_k(x_1), \ldots, \delta_k(x_n))$ for $k = 0, \ldots, K-1$. To find a proper model, we need a model selection strategy. In our situation, however, it is not appropriate to implement variable selection among all $nK + 1$ covariates associated with the parameters $1/E, \boldsymbol{\delta}_0(\boldsymbol{x}), \boldsymbol{\delta}_1(\boldsymbol{x}), \ldots, \boldsymbol{\delta}_{K-1}(\boldsymbol{x})$. Recall that $\delta_0(x)$ is interpreted as the initial bias effect and $\delta_k(x)$ is the systematic bias effect. It is thus more reasonable to keep each $\delta_k(x)$ as a whole parameter set in model selection. It can make the selected model more interpretable from the physical perspective.

Starting with the model including only the SSBM, we use the forward selection to add one $\delta_k(x)$ at a time. To begin with, we assume that the errors $\varepsilon(x_i, F_K)$ are independent with a normal distribution $\mathcal{N}(0, \sigma^2)$. Then the estimation of the parameters can be easily calculated by the maximum likelihood estimation (MLE). In each step of the forward selection, we select $\delta_k(x)$ that has the smallest root mean square error (RMSE) of the corresponding model, where

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{k=1}^{K}(v(x_i, F_k) - \hat{v}(x_i, F_k))^2}{df}}, \tag{73}$$

and *df* is the degrees of freedom in the corresponding model. Alternatively, we can use the Bayesian information criterion (BIC) to select $\delta_k(x)$ into the model at each step of the selection, where

$$\text{BIC} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K}(v(x_i, F_k) - \hat{v}(x_i, F_k))^2}{\sigma^2} + p \log N, \tag{74}$$

Here p is the number of parameters, and N is the number of observations in the corresponding regression model. If σ^2 in (74) is not available, an estimate $\hat{\sigma}^2$ can be obtained from the replicates. The R code for implementing the SPAR method is available from the authors upon request.

5.3.3 Example

In the image profiles of the nanobelt, the deflection is recorded at $n = 161$ points along the length of the nanobelt under $K = 15$ different applied forces. The length of NB is $L = 1252$ nm and the moment of inertia in the SSBM is $I = 8216510$ nm^4. Figure 24 shows the model selection results using the proposed method. The $\delta_k(x)$ is sequentially selected into the model in the following order: $\delta_0(x), \delta_{12}(x), \delta_{10}(x), \delta_8(x), \delta_9(x), \delta_6(x)$, and $\delta_2(x)$. It can be seen that after adding three or four terms, the decrease of RMSE starts to level off while the corresponding BIC value starts to increase. By considering both criteria, we take three ?k terms to build the final model. Thus, the chosen model

is

$$v(x, F_k) = \beta(x)F_k + \delta_0(x) + \delta_{10}(x)I(k > 10) + \delta_{12}(x)I(k > 12) + \varepsilon(x, F_k), \ \ k = 1, \ldots, K. \tag{75}$$

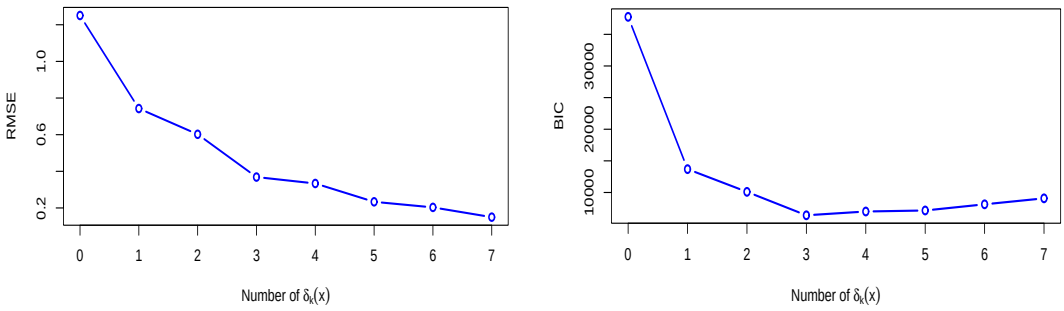
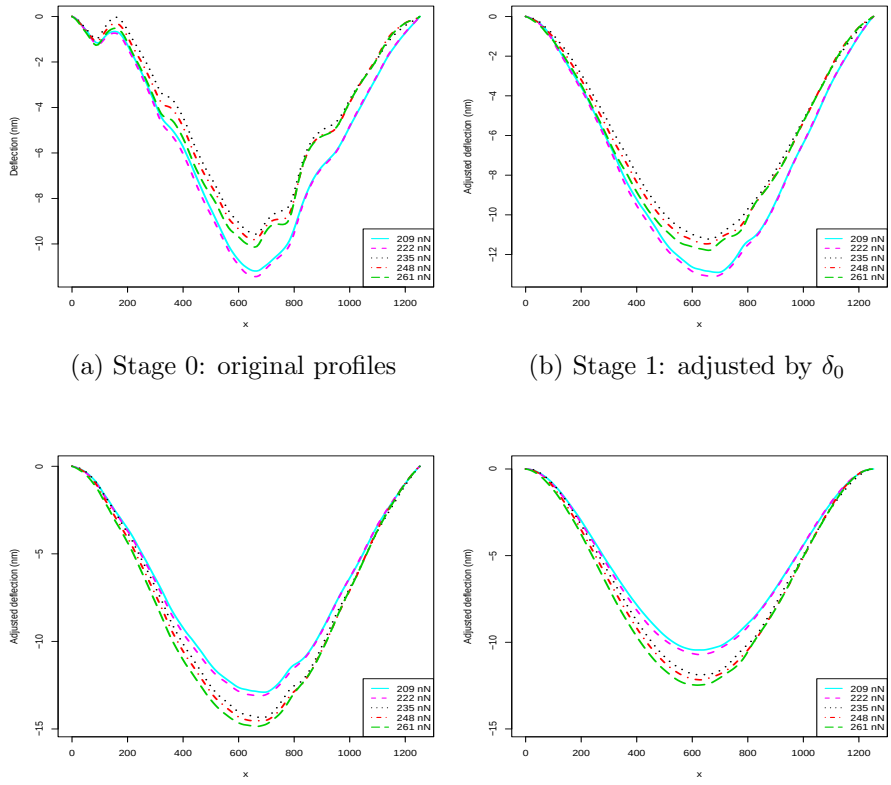


Figure 24: Forward model selection using RMSE and BIC on the NB data.

Here not only the initial bias $\delta_0(x)$ is significant, the systematic biases $\delta_{10}(x)$ and $\delta_{12}(x)$ also play an important role in modeling the data. To get more insights for the selected δ_k's in (75), at each stage of the selection, we define an adjusted deflection $v(x_i, F)_{adj}$ as $v(x_i, F)$ minus the selected δ_k's. For example, at stage 2, $v(x_i, F)_{adj} = v(x_i, F) - \delta_0(x) - \delta_{12}(x)I(k > 12)$. Note that the systematic bias $\delta_{12}(x)$ introduces the deflection into the image profiles starting from F_{13}, i.e., the profiles at $F = 235$, 248 and 261 nN. Similarly, $\delta_{10}(x)$ only brings in bias on the profiles under applied force F_{11} to F_{15}. Figure 25 shows the changes of five adjusted deflection profiles under applied forces F_{11} to F_{15} as the three δ_k terms are sequentially selected into the model.

The original five profiles are shown in Figure 25(a). When $\delta_0(x)$ is selected into the model at stage 1 of selection, it adjusts the initial bias among the five image profiles. In Figure 25(b), the adjusted deflection $v(xi, F)_{adj} = v(x_i, F) - \delta_0(x)$ looks closer to the SSBM, but the inconsistent pattern shown in Figure 21 still remains. Note that the inconsistent pattern appears between the profiles under $F_{11} = 209$ nN,



(a) Stage 0: original profiles

(b) Stage 1: adjusted by δ_0

(c) Stage 2: adjusted by δ_0 and δ_{12}

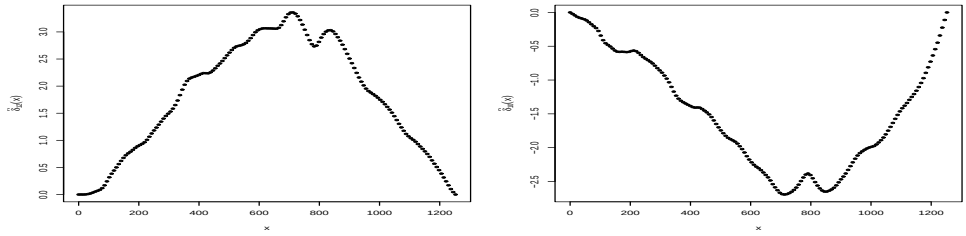
(d) Stage 3: adjusted by δ_0, δ_{12} and δ_{10}

Figure 25: Illustration of the adjusted deflection profiles under applied force from $F_{11} = 209$ nN to $F_{15} = 261$ nN.

$F_{12} = 222$ nN and those under $F_{13} = 235$ nN, $F_{14} = 248$ nN, $F_{15} = 261$ nN. At stage 2 of the selection, $\delta_{12}(x)$ is selected into the model. It further adjusts the profiles under the applied force F_{13}, F_{14}, and F_{15}. From Figure 25(c), we can see that the adjusted deflection $v(x_i, F)_{adj} = v(x_i, F) - \delta_0(x) - \delta_{12}(x)I(k > 12)$ is to push the profiles under the applied force F_{13}, F_{14}, and F_{15} to lie below those obtained at force F_{11}, and F_{12}. The inconsistency no longer exists in Figure 25(c). Therefore, adding $\delta_{12}(x)$ can remove the inconsistent pattern.

At stage 3 of the selection, $\delta_{10}(x)$ is chosen into the model. It can again adjust the five image profiles at the applied forces from $F_{11} = 209$ nN to $F_{15} = 261$ nN. As shown in Figure 25(b), to adjust the inconsistency among these five profiles, it is likely that the adjusted deflections have been pushed downwards too much. From Figure 25(d), we can see that adding $\delta_{10}(x)$ into the model is to pull all five profiles upwards and make the adjusted deflection $v(x_i, F)_{adj} = v(x_i, F) - \delta_0(x) - \delta_{12}(x)I(k > 12) - \delta_{10}(x)I(k > 10)$ better fit to the SSBM.

The two estimates $\hat{\delta}_{12}(x)$ and $\hat{\delta}_{10}(x)$ are shown in Figure 26. We can see that the opposite shapes of $\hat{\delta}_{12}(x)$ and $\hat{\delta}_{10}(x)$ in Figure 26 help remove the inconsistent pattern of the selected model and lead to a better fitting of the NB data.



(a) estimate of $\delta_{12}(x)$ (b) estimate of $\delta_{10}(x)$

Figure 26: Estimates of $\delta_{12}(x)$ and $\delta_{10}(x)$ from the selected model of NB.

We also compute R^2 to check the goodness-of-fit at each stage of model selection. The R^2 of fitting the SSBM is 85.88%, meaning that fitting the SSBM alone accounts for 85.88% of the total experimental variations. It shows that the SSBM fits the

data reasonably well in one statistical sense. However, the original curves (see Figure 21) do not look like the theoretical shape of the SSBM (see Figure 23). SPAR can identify and filter out, term by term, the observed deviations from the SSBM. The SSBM plus the initial bias term δ_0 fits the data better with $R^2 = 94.35\%$. The fit is further enhanced by adding two terms δ_{12} and δ_{10} with R^2 increased to 98.81%. The improvement due to the addition of these three terms is also evidenced from the profiles of the adjusted deflection $v(x, F)_{adj}$ based on the selected model shown in Figure 27. It is more consistent with the theoretical shape (see Figure 23) of the SSBM. Therefore, the selected model (75) can provide more reliable and precise estimation of the elastic modulus E.

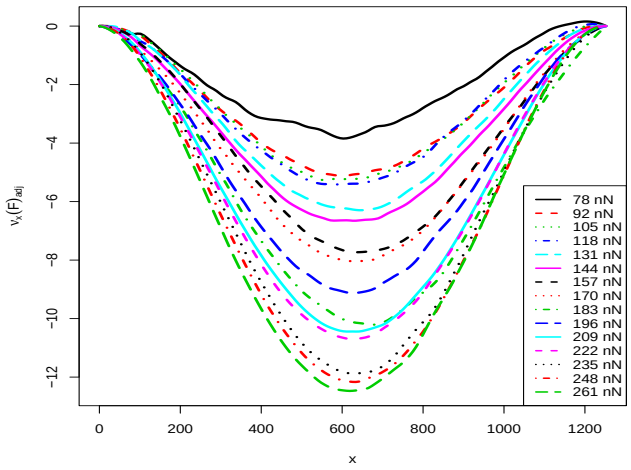
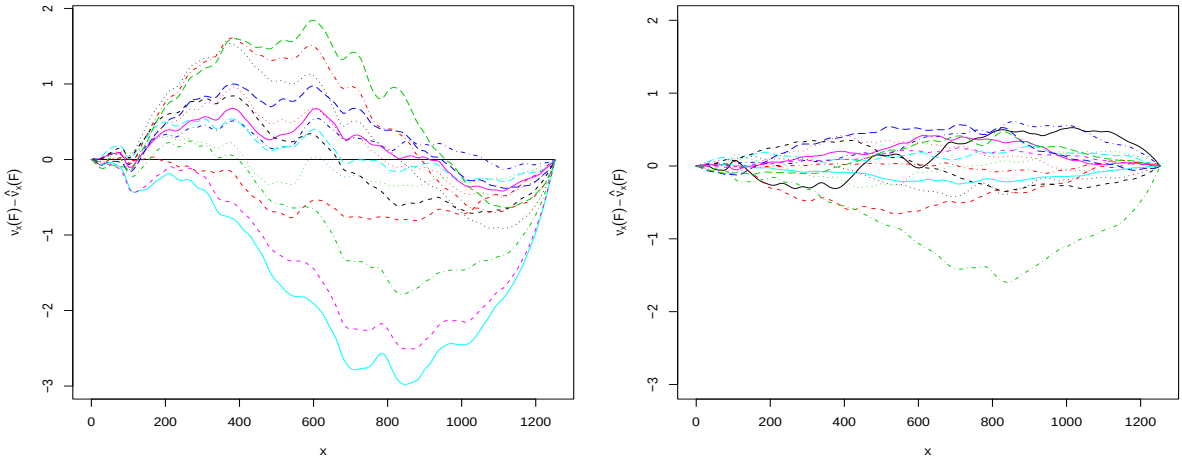


Figure 27: The image profiles for the adjusted deflection of NB.

To gauge the performance of the selected model using SPAR, we compare it with the MW method. The residual plots from these two approaches are shown in Figure 28. The residuals from the MW method show some systematic patterns, which indicates that the model needs improvement. No systematic pattern is observed in the residuals based on SPAR. Clearly, the selected model performs better. It removes the inconsistent pattern discussed above, while the MW method does not recognize this pattern. The residuals from the SPAR method are also much smaller.



(a) Residuals: MW method

(b) Residuals: proposed method

Figure 28: Comparison of two methods on the NB2 data.

Table 4 summarizes the estimation results using the two methods. Clearly SPAR gives a more precise estimate of the elastic modulus E. The standard error of E, $se(E)$, is reduced by 50%. The 95% confidence interval of E from SPAR is $(99.97, 103.07)$ and that from the MW method is $(91.24, 97.44)$. The non-overlapping of intervals suggests that one of the estimates can be misleading or wrong. Because SPAR incorporates the initial bias and adjusts the inconsistent pattern in the profiles, it is expected to provide more accurate determination of the elastic modulus than the MW method. To further verify this point, we perform SPAR using only half of the profiles of NB, i.e., the eight profiles under the applied force $F = 78, 105, 131, 157, 183, 209, 235, 261$ nN. The estimate of the elastic modulus GPa and the 95% confidence interval $(100.55, 104.79)$ are similar to those using SPAR with all the 15 profiles of NB. This shows that SPAR can give a more reliable estimate even with half of the profiles. Note that the confidence interval length for half profiles using SPAR is comparable to the corresponding length for full profiles using the MW method, thus confirming the 50% reduction in se(E).

Since the inconsistent pattern occurs for the last five image profiles, a simple

Table 4: Comparison of estimates with the NB data.

	$RMSE$	$1/\hat{E}$	std$(1/E)$	$\hat{E}$	std(E)
Mai and Wang	0.86	1.06e-02	1.77e-04	94.34	1.58
Proposed Method	0.37	9.85e-03	7.63e-05	101.52	0.79

alternative to SPAR, which the experimenter may favor, is to discard the last five profiles and apply the MW method to the first ten profiles. The resulting estimate of the elastic modulus is 96.23 GPa. The standard error is 1.51, which is almost twice as large as the standard error from applying SPAR to the 15 profiles. This shows that adjusting the inconsistent patterns and using the complete data is better than using only the profiles with consistent pattern for estimation.

5.4 Modeling with General Error Structures

The deflection of a nanobelt is a continuous and smooth phenomenon. For a given deflection curve, the error from the model at the distance $x = x_i$ should be positively correlated with those obtained near x_i, whereas the errors can be assumed to be independent between any two deflection curves.

As shown in the residual plots in Figure 28(b), the residuals have systematic patterns along the distance x. In particular, for a given force F_k, if the residual at x_i is large, then the residual at a distance close to x_i is likely to be large. It indicates that imposing some correlated error structure is warranted. Even though each profile curve was obtained by averaging 10 consecutive measurements to reduce the measurement error due to sources like equipment instability, there can be still errors after taking average. By incorporating this prior information and correlation structure, we can build a more general error structure as follows.

The model in (70) can be written as a linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where the response vector is $\boldsymbol{Y} = (v(x_1, F_1), \ldots, v(x_n, F_1), \ldots, v(x_1, F_K), \ldots, v(x_n, F_K))^T$, the parameters $\boldsymbol{\beta} = (1/E, \delta_0(x_1), \ldots, \delta_0(x_n), \delta_1(x_1), \ldots, \delta_1(x_n), \ldots, \delta_{K-1}(x_1), \ldots, \delta_{K-1}(x_n))^T$,

and $\boldsymbol{X}$ is the corresponding model matrix. Assuming that the error vector $\boldsymbol{\varepsilon}$ follows a normal distribution, i.e., $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, we consider the model Y = X?+? with a more general error structure $\boldsymbol{\Sigma}$, which consists of the K diagonal block submatrices

$$\boldsymbol{\Sigma}_k = \tau_k^2 \boldsymbol{R} + \sigma^2 \boldsymbol{I}, \quad k = 1, \ldots, K.$$

Here τ_k^2 is the error variance of the curve obtained at F_k, and $R = (r_{ij})_{n \times n}$ is the correlation matrix with r_{ij} quantifying the correlation between two deflections at the distance x_i and x_j obtained under the same applied force. The σ^2 term is used to quantify the error variation for each deflection curve obtained from averaging 10 consecutive measurements. The resulting covariance matrix $\boldsymbol{\Sigma}$ has a diagonal block structure since the errors between different deflection curves are considered to be independent. We use the Gaussian correlation function to model the correlation matrix R with

$$r_{ij} = \exp(-\theta(x_i - x_j)^2). \tag{76}$$

5.4.1 Parameter Estimation

The parameters in this model are $(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \theta)$, where $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_K^2)$. We consider the maximum likelihood estimation (MLE) for these parameters. For notational convenience, denote $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K)^T$, where $\boldsymbol{Y}_k \in \mathbb{R}^n$ as $\boldsymbol{Y}_k = (v(x_1, F_k), \ldots, v(x_n, F_k))^T$. Similarly, denote $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K)^T$, where $\boldsymbol{X}_k \in \mathbb{R}^{n \times p}$ is the part of the model matrix $\boldsymbol{X}$ corresponding to $\boldsymbol{Y}_k$, and p is the dimension of $\boldsymbol{\beta}$. The log-likelihood function for (**??**) can be written as

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \theta) &= -\frac{1}{2}\left[\log|\boldsymbol{\Sigma}| + (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\right] \\ &= -\frac{1}{2} \sum_{k=1}^{K} \left[\log|\boldsymbol{\Sigma}_k| + (\boldsymbol{Y}_k - \boldsymbol{X}_k\boldsymbol{\beta})^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{Y}_k - \boldsymbol{X}_k\boldsymbol{\beta})\right]. \end{aligned} \tag{77}$$

The last equality in (77) follows from applying the diagonal block structure of $\boldsymbol{\Sigma}$. The parameters $\boldsymbol{\tau}^2$ and θ are involved in every matrix inverse $\tilde{\boldsymbol{\Sigma}}_k^{-1}$. It requires

intensive computations to estimate these parameters by directly maximizing the log-likelihood function. Instead we use an algorithm to efficiently estimate parameters from $l(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \theta)$ by iteratively optimizing one while fixing the other two among $(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \theta)$ until convergence.

First observe that given $\boldsymbol{\tau}^2$ and θ, the MLE of $\boldsymbol{\beta}$ is its generalized least squares estimate

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}) \\ &= \left[\sum_{k=1}^{K}(\boldsymbol{X}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{X}_k)\right]^{-1}\left[\sum_{k=1}^{K}(\boldsymbol{X}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{Y}_k)\right]. \end{aligned} \tag{78}$$

For given $\boldsymbol{\beta}$ and θ, maximizing the log-likelihood function $l(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \theta)$ in (77) is equivalent to maximizing each $l_k(\boldsymbol{\beta}, \tau_k^2, \theta)$ individually, where

$$l_k(\boldsymbol{\beta}, \tau_k^2, \theta) = -\frac{1}{2}[\log|\boldsymbol{\Sigma}_k| + (\boldsymbol{Y}_k - \boldsymbol{X}_k\boldsymbol{\beta})^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{Y}_k - \boldsymbol{X}_k\boldsymbol{\beta})]. \tag{79}$$

Since τ_k^2 appears only in $l_k(\boldsymbol{\beta}, \tau_k^2, \theta)$ of the log-likelihood function (77), τ_k^2 can be individually estimated by maximizing $l_k(\boldsymbol{\beta}, \tau_k^2, \theta)$. Obviously, optimization with only one parameter is easy. For given $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, estimating θ is also a one-parameter optimization problem by maximizing the log-likelihood function $l(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \theta)$ in (77). Thus, if some initial estimates of $\boldsymbol{\beta}$ and θ are available, we can obtain the MLE of parameters through the following iterative algorithm:

Step 0. Obtain initial estimates $\hat{\theta}$ and $\hat{\boldsymbol{\beta}}$.

Step 1. Given $\hat{\theta}$ and $\hat{\boldsymbol{\beta}}$, update $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_K^2)$, i.e.,

$$\hat{\tau}_k^2 = \arg\min_{\tau_k^2}\left[\log|\boldsymbol{\Sigma}_k| + (\boldsymbol{Y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{Y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}})\right], \quad k = 1, \ldots, K. \tag{80}$$

Step 2. Given $\hat{\boldsymbol{\tau}}_2 = (\hat{\tau}_1^2, \ldots, \hat{\tau}_K^2)$ and $\hat{\boldsymbol{\beta}}$, update θ, i.e.,

$$\hat{\theta} = \arg\min_{\theta}\sum_{k=1}^{K}\left[\log|\boldsymbol{\Sigma}_k| + (\boldsymbol{Y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{Y}_k - \boldsymbol{X}_k\hat{\boldsymbol{\beta}})\right]. \tag{81}$$

Step 3. Given $\hat{\theta}$ and $\hat{\boldsymbol{\tau}}^2$, update $\boldsymbol{\beta}$, i.e.,

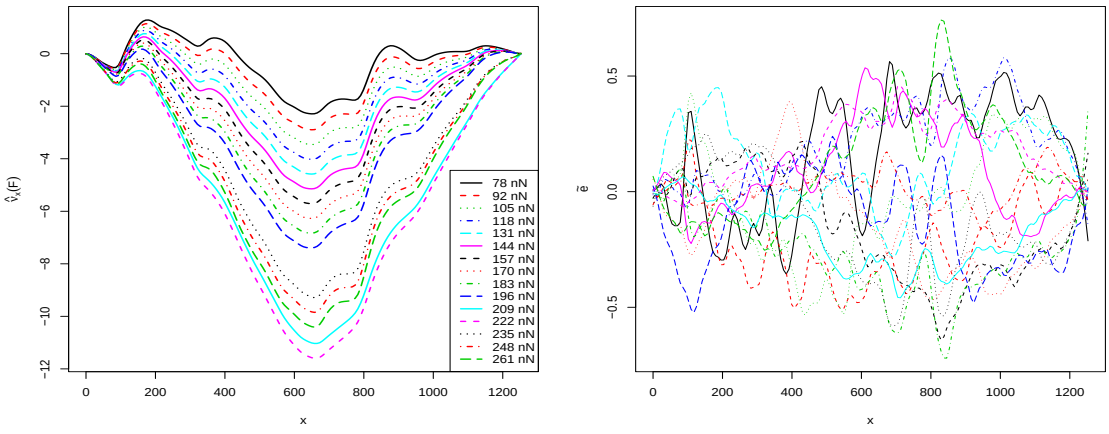
$$\hat{\boldsymbol{\beta}} = \left[\sum_{k=1}^{K}(\boldsymbol{X}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{X}_k)\right]^{-1}\left[\sum_{k=1}^{K}(\boldsymbol{X}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{Y}_k)\right]. \tag{82}$$

Step 4. Go to **Step 1** until convergence.

To obtain the initial estimates $\hat{\theta}$ and $\hat{\boldsymbol{\beta}}$, we use the ordinary least squares estimate for $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. Since θ is the parameter in the Gaussian correlation function in (76), we can take a relative large value as the initial estimate $\hat{\theta}$ (Santner et al., 2003).

5.4.2 Illustration

Now we apply the proposed general error structure $\boldsymbol{\Sigma}$ to the selected model (75). We define the generalized residual $\tilde{\boldsymbol{e}} = \hat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{Y} - \hat{\boldsymbol{Y}})$, where $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\Sigma}}$ are estimates of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$. Clearly, Figure 29(a) shows that the selected model fits the data well using the general error structure. Moreover, the generalized residuals in Figure (b) look much more random with the one in Figure 28(b). Therefore, the selected model is more appropriate than the one with independent error structure.



(a) fitting (b) generalized residual

Figure 29: Performance of the selected model using general error structure for NB.

To compare the efficiency of the estimates of E in the selected model with different error structures, we assume that the underlying model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. For convenience, denote E_{iid} as the elastic modulus parameter E in the model with independent and identically distributed (iid) error, and E_{gen} as the corresponding one in the model with general error structure. We obtain $\hat{E}_{gen} = 114.84$ GPa which is different from $\hat{E}_{iid} = 101.52$ GPa in Table 4. This difference can be explained by the fact that the residuals in Figure 28(b) have a clear pattern of positive correlation. The 95% confidence interval $(111.34, 118.34)$ of E_{gen} is disjoint with the 95% confidence interval $(99.97, 103.07)$ of E_{iid}. Although the length of the confidence interval of E_gen is larger, it is a more reasonable estimate because it incorporates the correlation structure in the deflection profiles. The accurate elastic modulus of ZnO nanowires is still under development (Bai et al., 2003, Song et al., 2005, Zhou et al., 2006) due to the practical difficulties mentioned before. Most people will agree that the elastic modulus of nanobelts on a 100 nm width scale should be close to that of bulk ZnO nanobelts ($140 \sim 180$ GPa) (Chen et al., 2006). In addition, the ZnO nanobelt analyzed in this paper has a smaller elastic modulus, compared to the other two nanobelts in Mai and Wang (2006). Thus, E_{gen} is expected to be a better estimate.

5.5 Discussions and Conclusions

In this article, we report a new method called SPAR to more precisely determine the elastic modulus of a nanobelt through statistical modeling and analysis of experimental data. It can automatically remove the initial bias, and adjust the systematic artifacts and errors introduced during measurement and thus can give a more precise and reliable estimate of the elastic modulus.

Due to the small size of nanomaterials, the noise from the uncertainty of complex boundary conditions, instrumental instability, and the measurement environment becomes relatively large compared to the actual scale of nanomaterials. It would be

difficult to conceive a physical model that can anticipate and incorporate all these sources of noise. Since the occurrence of these noises can vary from experiment to experiment, a catch-all model will be unwieldy for practical use. Statistical modeling is a more flexible and nimble alternative that can capture the noises that actually occur in an experiment. But a purely statistical approach lacks prediction power because the identified effects in one experiment may not carry over to another. On the other hand, a mechanics model with better physics can describe the intrinsic underlying properties and is thus more predictive. By avoiding the pitfalls of either approach, the proposed physical-empirical modeling approach can be a powerful tool. More discussions on this modeling and estimation technique can be found in Joseph and Melkote (2009) and the references therein.

The SPAR method is proposed and its performance studied for a specific experiment on nanobelts. It can, however, have broad applications in the quantification of the mechanical properties of 1D nanomaterials. For example, San Paulo et al. (2005) studied the mechanical elasticity of single and double clamped nanowires. The deflection of nanowires is measured by the controlled application of different normal forces with AFM. There is an initial variation in the growth of nanowires. Systematic bias can occur during the measurement under different applied forces. Therefore, SPAR can be used to get a better estimate of the elastic modulus. This new development demonstrates a statistical approach for quantifying the mechanical properties of 1D nanomaterials by comprehensively analyzing the acquired data and filtering out systematic artifacts.

The demonstrated methodology can be extended to other fields in nanotechnology. In the electrical measurements of nanodevices in a current range of pA (10^{-12} A), a precise identification of weak signals from the noise is essential for the reliable operation of chemical and biochemical sensors to avoid false alarms. For quantum devices and single electron transistors, the measured signal may be complicated by

instrumental instability and noise as well as measurement environment. In the application of piezoelectric nanowires for converting mechanical energy into electricity, the voltage generated from a nanowire depends on its dimension, the degree of its mechanical deformation, and the effectiveness of the charge output (Wang and Song, 2006). A statistical evaluation of the magnitude of the output voltage is essential for understanding the efficiency of the energy conversion. For all these applications, the demonstrated methodology can be effectively applied to filter out artifacts so that the operation of the devices can be more reliable and accurate. This research can serve as an example of a new cross-interdisciplinary effort between statistics and nanotechnology.

Materials and Methods

Materials The ZnO NBs were synthesized by a high temperature physical vapor deposition method inside a tube furnace (Pan et al., 2001). The NBs have a rectangular cross-section generally with 30-200 nm in width and thickness and 3-30 μm in length when controlling experimental parameters.

SEM imaging. A commercial scanning electron microscope (LEO 1530) was used to determine the morphology of ZnO NBs as well as the lateral dimensions of NBs and trenches.

AFM imaging and force measurement A commercial atomic force microscope (Asylum Research MFP3D) was used for imaging and force measurement. AFM image provided a reliable measurement of the thickness of the NBs. The force measurement was made by scanning the NB along its length direction using an AFM tip in contact mode at a constant applied force. A series of bending images of the NB were recorded by increasing the magnitude of the contact force. The AC240 cantilevers (spring constant of $\sim$ 2 N/m) from Asylum Research were used in our research, and each cantilever was carefully calibrated so that the AFM contact forces can be calculated.

APPENDIX A

EQUIVALENCE BETWEEN (9) AND (10)

From (8), $I(\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n, \boldsymbol{x}) = I(\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) + \kappa_x \boldsymbol{\eta}_x \boldsymbol{\eta}_x^T$, where $\kappa_x = e^{g(\boldsymbol{x})}/(1+e^{g(\boldsymbol{x})})^2$ and $\boldsymbol{\eta}_x = \frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{\theta}}$. Under mild regularity conditions, the Fisher information matrix $I(\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ is positive semi-definite and nonsingular. Therefore, applying the identity $\det(A + c\boldsymbol{x}\boldsymbol{x}^T) = \det(A)(1 + c\boldsymbol{x}^T A^{-1}\boldsymbol{x})$, we obtain

$$\begin{aligned}\det(I(\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n, \boldsymbol{x})) &= \det(I(\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) + \kappa_x \boldsymbol{\eta}_x \boldsymbol{\eta}_x^T) \\ &= \det(I(\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n))(1 + \kappa_x \boldsymbol{\eta}_x^T I^{-1}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)\boldsymbol{\eta}_x).\end{aligned}$$

Thus $\min_{\boldsymbol{x}} \det(I(\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n, \boldsymbol{x}))$ is the same as $\min_{\boldsymbol{x}} \kappa_x \boldsymbol{\eta}_x^T I^{-1}(\hat{\boldsymbol{\theta}}_n, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)\boldsymbol{\eta}_x$. Now under the constraint in (9), $\kappa_x = \alpha(1-\alpha)$ is a constant. Thus we get (10). Note that $\boldsymbol{\eta}_x = (-1/\sigma, -\log(\alpha/(1-\alpha))/\sigma, (x_1 - x_2)/\sigma)^T$ under constraint in (9). □

APPENDIX B

SOME PROOFS FOR SECTION 4.3.1

Lemma 2. *Let A and B be $n \times n$ nonsingular matrices. If $A - B$ is nonsingular, then $B^{-1} - A^{-1}$ is nonsingular. Further,*

$$(B^{-1} - A^{-1})^{-1} = B + B(A - B)^{-1}B. \tag{83}$$

Proof: Clearly,

$$B^{-1} - A^{-1} = (I - A^{-1}B)B^{-1} \tag{84}$$

$$A - B = A(I - A^{-1}B) \tag{85}$$

Hence,

$$(A - B)^{-1} = (I - A^{-1}B)^{-1}A^{-1} \tag{86}$$

Using the above equations, then,

$$\begin{aligned}
&(B^{-1} - A^{-1})(B + B(A - B)^{-1}B) \\
&\qquad = (I - A^{-1}B)B^{-1}(B + B(A - B)^{-1}B) \\
&\qquad = (I - A^{-1}B)(I + (A - B)^{-1}B) \\
&\qquad = (I - A^{-1}B)(I + (I - A^{-1}B)^{-1}A^{-1}B) \\
&\qquad = I - A^{-1}B + A^{-1}B \\
&\qquad = I
\end{aligned}$$

Similarly, it is easy to verify $(B + B(A - B)^{-1}B)(B^{-1} - A^{-1}) = I$. So,

$$(B^{-1} - A^{-1})^{-1} = B + B(A - B)^{-1}B. \qquad \square$$

Denote $A \succ 0$ as A is a positive matrix. Define $\|A\|_2$ as the L_2 norm of matrix A.

Lemma 3. *A and B are $n \times n$ real symmetric matrices. Suppose $A \succ 0$, $B \succ 0$, and $A - B \succ 0$, then,*

(1). $B^{-1} - A^{-1} \succ 0$

(2). $\|A + B\|_2 > \|A\|_2 > \|B\|_2$

(3). $\|B^{-1}\|_2 > \|A^{-1}\|_2$

(4). $\|BA^{-1}B\|_2 \leq \|B\|_2$

Proof:

(1). By Lemma 2, we have $(B^{-1} - A^{-1})^{-1} = B + B(A - B)^{-1}B$. Since $B \succ 0$ and $A - B \succ 0$,

$$A - B \succ 0 \Rightarrow (A - B)^{-1} \succ 0 \Rightarrow B(A - B)^{-1}B \succ 0$$
$$\Rightarrow B + B(A - B)^{-1}B \text{ is positive definite}$$
$$\Rightarrow (B^{-1} - A^{-1})^{-1} \text{ is positive definite}$$
$$\Rightarrow B^{-1} - A^{-1} \text{ is positive definite}$$

That is $B^{-1} - A^{-1} \succ 0$. $\square$

(2). It is known that if A is real, symmetric, and positive definite, then,

$$\|A\|_2 = \lambda_{\max}(A) = \max_{\|x\|_2 \leq 1} x^T A x,$$

where $\lambda_{\max}(A)$ is defined as he largest eigenvalue of A. Knowing both A and B are positive definite, obviously $A + B$ is also positive definite. Then

$$\|A + B\|_2 = \lambda_{\max}(A + B) = \max_{\|x\|_2 \leq 1} x^T (A + B) x$$
$$> \max_{\|x\|_2 \leq 1} x^T A x = \lambda_{\max}(A) = \|A\|_2.$$

So $\|A + B\|_2 > \|A\|_2$. Since B and $A - B$ are positive definite, we have,

$$\|A\|_2 = \|B + (A - B)\|_2 > \|B\|_2.$$

Therefore, $\|A+B\|_2 > \|A\|_2 > \|B\|_2$. $\square$

(3). Using the result in (1), we have $B^{-1} - A^{-1} \succ 0$. Applying it into the result of (2), clearly,

$$\|B^{-1}\|_2 = \|A^{-1} + (B^{-1} - A^{-1})\|_2 > \|A^{-1}\|_2.$$

That is $\|B^{-1}\|_2 > \|A^{-1}\|_2$. $\square$

(4). From $A - B \succ 0$, we have $B^{-1} - A^{-1} \succ 0$, for $y \in \mathbb{R}^n$,

$$y^T(B^{-1} - A^{-1})y \geq 0.$$

That is $y^T B^{-1} y \geq y^T A^{-1} y$. $\forall x \in \mathbb{R}^n$, let $y = Bx$, then,

$$(x^T B)B^{-1}(Bx) \geq (x^T B)A^{-1}(Bx).$$

That is $(x^T Bx) \geq x^T BA^{-1}Bx$. So,

$$\max_{\|x\|_2 \leq 1} x^T Bx \geq \max_{\|x\|_2 \leq 1} x^T BA^{-1}Bx.$$

It is to say $\|B\|_2 \geq \|BA^{-1}B\|_2$. One more step, we have

$$\|BA^{-1}B\|_2 \leq \|B\|_2 \leq \|A\|_2 \quad \square$$

Remark: If we think A and B as operators, assuming A, B, and $A-B$ are positive operators, then Lemma 3 also holds.

Proof of Lemma 1: For any given i, from assumption, $\|p_i^{-1}\| \sim K$ and $\|p_i\| \sim 1/K$. Since $(p_i + \lambda w) - p_i \succ 0$, from Lemma 3, we have $\|(p_i + \lambda w)^{-1}\| \leq \|p_i^{-1}\|$. Then,

$$\begin{aligned}
\|(p_i + \lambda w)^{-1} p_i\| &\leq \|(p_i + \lambda w)^{-1}\| \|p_i\| \\
&\leq \|p_i^{-1}\| \|p_i\| \\
&\leq 1.
\end{aligned}$$

Hence, we can get

$$\begin{aligned}\|(p_i+\lambda w-p_i^2)^{-1}(p_i+\lambda w)\| &= \|(\mathbf{1}-(p_i+\lambda w)^{-1}p_i^2)^{-1}\| \\ &\le \frac{1}{1-\|(p_i+\lambda w)^{-1}p_i^2\|} \\ &\le \frac{1}{1-\|(p_i+\lambda w)^{-1}p_i\|\|p_i\|} \\ &\le \frac{1}{1-\|p_i\|}\ .\end{aligned}$$

The first inequality is based on the condition that $\|(p_i+\lambda w)^{-1}p_i^2\| < 1$ which is provided from $\|(p_i+\lambda w)^{-1}p_i\| \le 1$.

From the proof of **Lemma 1**, we know that $\|(p_i+\lambda w)^{-1}p_i^2\| < \|p_i\|$ which is provided from $\|(p_i+\lambda w)^{-1}p_i\| \le 1$. Then

$$\begin{aligned}\|(p_i+\lambda w_i)^{-1}(p_i+\lambda w_i-p_i^2)\|_2 &\le \|1-(p_i+\lambda w_i)p_i^2\|_2 \\ &\le 1+\|(p_i+\lambda w_i)p_i^2\|_2 \\ &\le 1+\|p_i\|_2. \quad \square\end{aligned}$$

Lemma 4. *Suppose p_i, $p_i-p_i^2$, and w_i are positive definite. If $\forall i \in \{1,2,\ldots,K-1\}$, $\|p_i\|_2 \sim 1/K$ and $\|p_i^{-1}\|_2 \sim K$, then,*

$$\|I+A^{-1}P(\mathbf{1}-P^TA^{-1}P)^{-1}P^T\|_2 \le O(K),$$

where operator matrix $A = diag(p_1+\lambda w_1,\ldots,p_{K-1}+\lambda w_{K-1})$ and operator vector $P=(p_1,p_2,\ldots,p_{K-1})^T$.

Proof: Obviously,

$$\begin{aligned}&\|I+A^{-1}P(\mathbf{1}-P^TA^{-1}P)^{-1}P^T\|_2 \\ &\le 1+\|A^{-1}P(\mathbf{1}-P^TA^{-1}P)^{-1}P^T\|_2 \\ &\le 1+\|A^{-1}P\|_2\|(\mathbf{1}-P^TA^{-1}P)^{-1}\|_2\|P\|_2 \qquad (87)\end{aligned}$$

From $(\mathbf{1} - P^T A^{-1} P)^{-1} = \sum_{j=1}^{\infty} (P^T A^{-1} P)^j$, one can have

$$\begin{aligned} \|(\mathbf{1} - P^T A^{-1} P)^{-1}\|_2 &\leq \sum_{j=0}^{\infty} \|P^T A^{-1} P\|_2^j = \frac{1}{1 - \|P^T A^{-1} P\|_2} \\ &\leq \frac{1}{1 - \sum_{i=1}^{K-1} \|p_i (p_i + \lambda w_i)^{-1} p_i\|_2} \\ &\leq \frac{1}{1 - \sum_{i=1}^{K-1} \|p_i\|_2} \end{aligned} \tag{88}$$

The last inequality utilizes the result from (4) of Lemma 3, provided that $p_i \succ 0$, $p_i + \lambda w_i \succ 0$, and $(p_i + \lambda w_i) - p_i \succ 0$.

From the proof of Lemma 2.1, it is known that $\forall i$, $\|(p_i + \lambda w_i)^{-1} p_i\|_2 \leq 1$, then

$$\|A^{-1} P\|_2 \leq \sqrt{\sum_{i=1}^{K-1} \|(p_i + \lambda w_i)^{-1} p_i\|_2^2} \leq \sqrt{K-1}.$$

Similarly, we have $\|P\|_2 \leq \sqrt{\sum_{i=1}^{K-1} \|p_i\|_2^2}$. Therefore,

$$\begin{aligned} &\|I + A^{-1} P (\mathbf{1} - P^T A^{-1} P)^{-1} P^T\|_2 \\ &\leq 1 + \|A^{-1} P (\mathbf{1} - P^T A^{-1} P)^{-1} P^T\|_2 \\ &\leq 1 + \frac{\|A^{-1} P\|_2 \|P\|_2}{1 - \|P^T A^{-1} P\|_2} \\ &\leq 1 + \frac{\sqrt{K-1} \sqrt{\sum_{i=1}^{K-1} \|p_i\|_2^2}}{1 - \sum_{i=1}^{K-1} \|p_i\|_2} \end{aligned}$$

If assuming $\forall p_i, i \in \{1, 2, \ldots, K-1\}$, $\|p_i\|_2 \sim 1/K$, then

$$\begin{aligned} \|I + A^{-1} P (\mathbf{1} - P^T A^{-1} P)^{-1} P^T\|_2 &\leq 1 + \frac{\sqrt{K-1} \sqrt{(K-1)/K^2}}{1 - (K-1)/K} \\ &\approx O(K) \quad \square \end{aligned}$$

Lemma 5. *(Lemma 2.2, Cox and O'Sullivan, 1990) For $f \in N_{f^0}$, $b > 0$ and $v = 1, 2, \ldots$, suppose $\gamma_\nu \approx \nu^r$ for some $r > 0$, meaning γ_ν / ν^r is bounded away from 0 and ∞ as $\nu \to \infty$. Then for $b \geq 0$ and $c \geq 0$ with $b + c < 2 - 1/r$, uniformly in $f \in N_{f^0}$,*

$$\sum_{\nu} (1 + \gamma_\nu^b)(1 + \gamma_\nu^c)(1 + \lambda \gamma_\nu)^{-2} \approx \lambda^{-(b+c+1/r)} \quad \text{as } \lambda \to 0, \tag{89}$$

and for any given operator $H(f)$,

$$\|(u+\lambda w)^{-1}(f)H(f)\zeta\|_b^2 \approx \sum_{\nu}(1+\gamma_\nu^b)(1+\lambda\gamma_\nu)^{-2}\langle H(f)\zeta,\phi_\nu\rangle^2 \tag{90}$$

Lemma 6. *Notation follows Cox and O'Sullivan (1990). Suppose we have*

$$\|(U+\lambda W)^{-1}\lambda W f\|_b = \sum_{\nu=1}\frac{(\lambda\gamma_\nu)^2}{(1+\lambda\gamma_\nu)^2}(1+\gamma_\nu)^b\langle f,U\phi_\nu\rangle \tag{91}$$

where U *and* W *are operators and* $\gamma_\nu = \nu^r,\ \nu = 1,2,\ldots$ *for some* r. $\{\phi_\nu\}$ *is an orthogonal basis. Then for* $0 \le b \le p \le 1$,

$$\|(U+\lambda W)^{-1}\lambda W f\|_b^2 \approx M\lambda^{p-b}, \tag{92}$$

where $\|f\|_b = \sum_{\nu=1}(1+\gamma_\nu)^b\langle f,U\phi_\nu\rangle^2$.

Proof: Obviously, $\gamma_\nu < 1+\gamma_\nu < 2\gamma_\nu$ and $\lambda > \lambda^2$ when $\lambda < 1$. Then

$$\begin{aligned}
\|(U+\lambda W)^{-1}\lambda W f\|_b^2 &= \sum_{\nu=1}\frac{(\lambda\gamma_\nu)^2}{(1+\lambda\gamma_\nu)^2}\frac{(1+\gamma_\nu)^b}{(1+\gamma_\nu)^p}(1+\gamma_\nu)^p\langle f,U\phi_\nu\rangle^2 \\
&\le \sum_{\nu=1}\frac{(\lambda\gamma_\nu)^2}{(1+\lambda\gamma_\nu)^2}\frac{(1+\gamma_\nu)^b}{(1+\gamma_\nu)^b\gamma_\nu^{p-b}}(1+\gamma_\nu)^p\langle f,U\phi_\nu\rangle^2 \\
&= \sum_{\nu=1}\frac{\lambda^{p-b}(\lambda\gamma_\nu)^{2-p+b}}{(1+\lambda\gamma_\nu)^2}(1+\gamma_\nu)^p\langle f,U\phi_\nu\rangle^2 \\
&\le \lambda^{p-b}\sum_{\nu=1}(1+\gamma_\nu)^p\langle f,U\phi_\nu\rangle^2 \\
&= \lambda^{p-b}\|f\|_p^2,
\end{aligned} \tag{93}$$

where $0 \le b < p \le 1$. This inequality is (4.9) in Cox and O'Sullivan (1990). The result in (93) is an upper bound. We do the following derivation to get a lower bound.

$$\begin{aligned}
&\sum_{\nu=1}\frac{(\lambda\gamma_\nu)^2}{(1+\lambda\gamma_\nu)^2}\frac{(1+\gamma_\nu)^b}{(1+\gamma_\nu)^p}(1+\gamma_\nu)^p\langle f,U\phi_\nu\rangle^2 \\
&\ge \sum_{\nu=1}\frac{1}{2^p}\frac{(\lambda\gamma_\nu)^2}{(1+\lambda\gamma_\nu)^2}\frac{\gamma_\nu^b}{\gamma_\nu^p}(1+\gamma_\nu)^p\langle f,U\phi_\nu\rangle^2 \\
&= \sum_{\nu=1}\frac{1}{2^p}\frac{\lambda^2\gamma_\nu^{2-p+b}}{(1+\lambda\gamma_\nu)^2}(1+\gamma_\nu)^{p-b}(1+\gamma_\nu)^b\langle f,U\phi_\nu\rangle^2 \\
&= \frac{\lambda^{p-b}}{2^p}\sum_{\nu=1}\frac{\lambda^{2-p+b}\gamma_\nu^{2-p+b}(1+\gamma_\nu)^{p-b}}{(1+\lambda\gamma_\nu)^2}(1+\gamma_\nu)^b\langle f,U\phi_\nu\rangle^2.
\end{aligned} \tag{94}$$

Since $\lambda^{2-p+b} > \lambda^2$ when $\lambda < 1$, $\frac{\lambda^{2-p+b}\gamma_\nu^{2-p+b}(1+\gamma_\nu)^{p-b}}{(1+\lambda\gamma_\nu)^2}$ can be bounded below for large γ_ν. Therefore,

$$\begin{aligned}
&\sum_{\nu=1} \frac{(\lambda\gamma_\nu)^2}{(1+\lambda\gamma_\nu)^2} \frac{(1+\gamma_\nu)^b}{(1+\gamma_\nu)^p} (1+\gamma_\nu)^p \langle f, U\phi_\nu\rangle^2 \\
&\geq \frac{c}{2^p} \lambda^{p-b} \sum_{\nu=1} (1+\gamma_\nu)^b \langle f, U\phi_\nu\rangle^2 \\
&= c_1 \lambda^{p-b} \|f\|_b^2 \qquad (95)
\end{aligned}$$

Combining the results in (93) and (95), we declare that

$$\|(U+\lambda W)^{-1}\lambda W f\|_b^2 \approx M\lambda^{p-b}. \qquad (96)$$

□

Lemma 7. *There are two self-adjoint operators U and W. Suppose $\|U\| \sim 1/K$ and we use the simultaneous diagonalization as*

$$\langle KU\phi_\nu, \phi_\nu\rangle = 1, \quad \langle W\phi_\nu, \phi_\nu\rangle = \gamma_\nu. \qquad (97)$$

where $\{\phi_\nu\}$ is an orthogonal basis. Assume $\gamma_\nu = \nu^r$ for some $r > 0$, and $\nu = 1, 2, \ldots$, Define an operator B as $B = (U+\lambda W)^{-1}\lambda W$, then

$$\|Bx\|_b^2 \approx M(\lambda K)^{p-b} \quad \text{as } \lambda \to 0. \qquad (98)$$

Proof: From the definition of the simultaneous diagonalization (Rao, 1973), we know that

$$W\phi_\nu = \gamma_\nu KU\phi_\nu, \quad x = \sum_\nu \langle x, KU\phi_\nu\rangle \phi_\nu; \qquad (99)$$

Note that U and W are self-adjoint operators, i.e., $\langle Ux, y\rangle = \langle x, Uy\rangle$. Based on (97) and (99), we have

$$(U+\lambda W)\phi_\nu = (\frac{1}{K} + \lambda\gamma_\nu) KU\phi_\nu, \qquad (100)$$

and

$$\begin{aligned}\langle (U+\lambda W)x, \phi_\nu\rangle &= \langle x, (U+\lambda W)\phi_\nu\rangle \\ &= (1+\lambda K\gamma_\nu)\langle x, U\phi_\nu\rangle \\ &= (\frac{1}{K}+\lambda\gamma_\nu)\langle x, KU\phi_\nu\rangle. \end{aligned} \tag{101}$$

Hence,

$$(U+\lambda W)^{-1}KU\phi_\nu = (\frac{1}{K}+\lambda\gamma_\nu)^{-1}\phi_\nu. \tag{102}$$

This implies that

$$\begin{aligned}(U+\lambda W)^{-1}Ux &= \sum_\nu \langle (U+\lambda W)^{-1}Ux, KU\phi_\nu\rangle\phi_\nu \\ &= \sum_\nu \langle Ux, (U+\lambda W)^{-1}KU\phi_\nu\rangle\phi_\nu \\ &= \sum_\nu (\frac{1}{K}+\lambda\gamma_\nu)^{-1}\langle Ux, \phi_\nu\rangle\phi_\nu \\ &= \sum_\nu \frac{1}{K}(\frac{1}{K}+\lambda\gamma_\nu)^{-1}\langle x, KU\phi_\nu\rangle\phi_\nu \\ &= \sum_\nu (1+\lambda K\gamma_\nu)^{-1}\langle x, KU\phi_\nu\rangle\phi_\nu. \end{aligned} \tag{103}$$

Obviously, $B = I - (U+\lambda W)^{-1}U$, then,

$$\begin{aligned}Bx &= \sum_\nu \langle x, KU\phi_\nu\rangle\phi_\nu - \sum_\nu (1+\lambda K\gamma_\nu)^{-1}\langle x, KU\phi_\nu\rangle\phi_\nu \\ &= \sum_\nu [1-(1+\lambda K\gamma_\nu)^{-1}]\langle x, KU\phi_\nu\rangle\phi_\nu \\ &= \sum_\nu \frac{\lambda K\gamma_\nu}{1+\lambda K\gamma_\nu}\langle x, KU\phi_\nu\rangle\phi_\nu \end{aligned} \tag{104}$$

Since $(KU+W)\phi_\nu = (1+\gamma_\nu)KU\phi_\nu$, the intermediate space between U and W can be extended as

$$\langle \theta, \xi\rangle_b = \sum_\nu (1+\gamma_\nu)^b \langle \theta, KU\phi_\nu\rangle\langle \xi, KU\phi_\nu\rangle \tag{105}$$

Then

$$\|Bx\|_b^2 = \sum_\nu [\lambda K\gamma_\nu/(1+\lambda K\gamma_\nu)]^2(1+\gamma_\nu)^b\langle x, KU\phi_\nu\rangle\phi_\nu \tag{106}$$

Using the result (96) in Lemma 6, we have

$$\|Bx\|_b^2 \approx M(\lambda K)^{p-b} \quad \text{as } \lambda \to 0. \tag{107}$$

□

Proof of Proposition 1: Note that in the case of $p = 1$ and $b = 0$, which corresponds to the usual integrated squared error, we consider $p_i(1-p_i)$ as U in Lemma 7 , and the penalty term w as W in Lemma 7. From the assumption $\|p_i(1-p_i)\| \sim 1/K$, the result follows immediately by applying Lemma 7. □

Lemma 8. *Notation follows the same as Lemma 6, Define an operator* $G = U + \lambda W$. *If there is a function* $Dl_n(\theta) - Dl(\theta)$ *such that*

$$E[\{Dl_n(\theta) - Dl(\theta)\}\phi_\nu] \sim \frac{1}{nK},$$

then in the case of $b = 0$,

$$\|G^{-1}[Dl_n(\theta) - Dl(\theta)]\|_b^2 = (nK)^{-1}\lambda^{-1/2m}K^{(2-1/2m)} \quad \textit{as } \lambda \to 0. \tag{108}$$

Proof: First, let us check some calculations on the inverse of operator G,

$$\begin{aligned} G^{-1}x &= \sum_\nu \langle G^{-1}x, KU\phi_\nu\rangle\phi_\nu \\ &= \sum_\nu \langle x, (U+\lambda W)^{-1}KU\phi_\nu\rangle\phi_\nu \\ &= \sum_\nu \langle x, (\frac{1}{K} + \lambda K\gamma_\nu)^{-1}\phi_\nu\rangle\phi_\nu \\ &= \sum_\nu K(1+\lambda K\gamma_\nu)^{-1}\langle x, \phi_\nu\rangle\phi_\nu \end{aligned} \tag{109}$$

Then in the case of $b = 0$, which corresponds to the usual integrated squared error, recall Lemma 3.4 and we get

$$\begin{aligned} \|G^{-1}[Dl_n(\theta) - Dl(\theta)]\|_b^2 &= \sum_\nu K^2(1+\lambda K\gamma_\nu)^{-2}[\{Dl_n(\theta) - Dl(\theta)\}\phi_\nu]^2 \\ &\approx K^2(nK)^{-1}(\lambda K)^{-1/2m} \\ &= (nK)^{-1}\lambda^{-1/2m}K^{(2-1/2m)} \quad \text{as } \lambda \to 0. \end{aligned} \tag{110}$$

□

APPENDIX C

SOME PROOFS FOR SECTIONS 4.3.2 AND 4.3.3

We denote $\mathcal{EG}(M)$ as the set of all eigenvalues of M. For an eigenvalue $\lambda_r(M) \in \mathcal{EG}(M)$, we denote $\gamma_r(M)$ as the corresponding eigenvector. The $\mathcal{EV}(M, \eta_r(M))$ denotes the set of all eigenvectors of M for $\eta_r(M)$. Lemma 9 – Lemma 11 haven not contained the proof. Detail on the proofs can be found in Kneip (1994).

Lemma 9. *Let M^0 and M^* be real $n \times n$ matrices, $n \in \mathbb{N}$. Suppose that M^0 is symmetric and that, for some n_0,$1 \leq n_0 \leq n$, it holds that $\eta_1(M^0) \geq \eta_2(M^0), \cdots, \eta_{n_0}(M^0) \geq \eta_{n_0+1}(M^0)$ and $\eta_{n_0+1}(M^0) = \cdots = \eta_n(M^0) = 1$.*

For $\eta \in \mathcal{EG}(M^0)$ use $U(\eta)$ to denote the projection matrix projecting onto the eigenspace of M^0 for η, and let

$$\boldsymbol{S}(\eta) = \sum_{\tau \in \mathcal{EG} \setminus \{\eta\}} \frac{1}{\tau - \eta} U(\tau)$$

be the reduced resolvent of M^0 for η. Furthermore, for $\beta > 0$ let $\mathcal{EG}^*(M^0, M^*, \beta)$ denote the set of all $\eta \in \mathcal{EG}(M^0)$ such that

$$\|\bar{\boldsymbol{S}}(\eta)^{(p_1)} M^* \bar{\boldsymbol{S}}(\eta)^{(p_2)} \cdots \bar{\boldsymbol{S}}(\eta)^{(p_q)} M^* \bar{\boldsymbol{S}}(\eta)^{(p_{q+1})}\| \leq \frac{\beta^q}{\|\boldsymbol{S}(\eta)\|^{q-p}} \tag{111}$$

holds for all $q \in \mathbb{N}$ and $p_1, \ldots, p_{q+1} \in \mathbb{N} \cup \{0\}$ with $p_1 + \ldots + p_{q+1} = p \leq q$. Here, $\bar{\boldsymbol{S}}(\eta)^{(0)} = -U(\eta)$ and $\bar{\boldsymbol{S}}(\eta)^{(\delta)} = \boldsymbol{S}(\eta)^\delta$, $\delta \in \mathbb{N}$.

Then if, for some $\beta < 1/8$, $\eta_r(M^0) \in \mathcal{EG}^*(M^0, M^*, \beta)$ holds for all $r = 1, \ldots, n_0$, then for any r, there is a real $\tau_r \in \mathcal{EG}(M^0 + M^*)$ and a real eigenvector $u_r \in \mathcal{EV}(M^0 + M^*, \tau_r)$ such that

$$\left|\tau_r - \eta_r(M^0) - tr(U(\eta_r(M^0)) M^* U(\eta_r(M^0)))\right| \leq \alpha(\eta_r(M^0))_1 \times \frac{1}{1 - 4\beta}, \tag{112}$$

$$\|\gamma_r(M^0) - u_r\|_2 \leq \alpha(\eta_r(M^0))_2 \times \frac{2}{1 - 4\beta}, \tag{113}$$

where

$$\alpha(\eta_r(M^0))_1 = \sup_{q\geq 2} \sup_{p_1+\ldots+p_q=q-1} \frac{|tr(M^*\bar{\boldsymbol{S}}(\eta)^{(p_1)}\cdots M^*\bar{\boldsymbol{S}}(\eta)^{(p_q)})|}{\beta^{q-2}},$$
$$\alpha(\eta_r(M^0))_2 = \sup_{q\geq 1} \sup_{p_1+\ldots+p_q=q} \frac{\|\bar{\boldsymbol{S}}(\eta)^{(p_1)}M^*\cdots\bar{\boldsymbol{S}}(\eta)^{(p_q)}M^*U(\eta)\|}{\beta^{q-1}}.$$

Remark: The Lemma 9 is too general to be feasible for practical computation. There are several different ways to make them more specific and useful. The following lemma provides a modified version which is suitable for the proof of Theorem 2.

Lemma 10. *Under the conditions of Lemma 9, define the matrices* $\bar{\boldsymbol{S}}(\eta)$ *by replacing* $\tau-\eta$ *by* $|\tau-\eta|$ *in* (9). *For* $\eta\in\mathcal{EG}(M^0)$, *set*

$$\beta(\eta) = \max\{\|M^*\cdot S(\eta)\|, \|S(\eta)\|\cdot\|M^*\cdot U(\eta)\|\},$$

and let $\beta = \max_{r=1,\ldots,n_0}\beta(\eta_r(M^0))$, *then* $\eta_r(M^0)\in\mathcal{EG}^*(M^0,M^*,\beta)$, *for all* $r = 1,\ldots,n_0$, *and*

$$\alpha(\eta_r(M^0))_1 \leq \|M^*U(\eta_r(M^0))\|\cdot\|U(\eta_r(M^0))M^*S(\eta_r(M^0))\|;$$
$$\alpha(\eta_r(M^0))_2 \leq \|S(\eta_r(M^0))\|\cdot\|M^*U(\eta_r(M^0))\|.$$

The following lemma is particularly suited for the proof of Theorem 3.

Lemma 11. *Under the condition of Lemma 9, suppose that* M^* *is symmetric. Define matrices* $\bar{\boldsymbol{S}}(\eta)$ *by replacing* $\tau-\eta$ *by* $|\tau-\eta|$ *in* (9). *For* $\eta\in\mathcal{EG}(M^0)$, *set* $s(\eta) = \|S(\eta)\|$,

$$\beta(\eta) = \max\{\|U(\eta)M^*U(\eta)\|\cdot s(\eta), \|S(\eta)^{1/2}M^*U(\eta)\|\cdot s(\eta)^{1/2}\|S(\eta)^{1/2}M^*S(\eta)^{1/2}|\},$$

and let $\beta = \max_{\eta\in\mathcal{EG}(M^0)}\beta(\eta)$. *Then* $\mathcal{EG}^*(M^0,M^*,\beta) = \mathcal{EG}(M^0)$, *and, for any* $\eta\in\mathcal{EG}(M^0)$,

$$\alpha(\eta_r(M^0))_1 \leq s(\eta)|tr(U(\eta)M^*U(\eta)M^*U(\eta))| + |tr(U(\eta)M^*S(\eta)M^*U(\eta))|,$$
$$\alpha(\eta_r(M^0))_2 \leq s(\eta_r(M^0))^{1/2}\cdot\|S(\eta_r(M^0))^{1/2}M^*U(\eta_r(M^0))\|.$$

Lemma 12. *Under the normal assumptions, i.e., the notation and assumption are the same as what mentioned in the first part. Define* $H = A_1^T(\lambda) + A_1(\lambda) + A_1^T(\lambda)A_1(\lambda)$, *where* $A_1(\lambda) = -G_\lambda^{-1}\lambda W$. *Then we can have*

$$\|H\|_2^2 \sim O(\lambda K). \tag{114}$$

Proof: In the definition of H, the term $A_1^T(\lambda)A_1(\lambda)$ has a higher order than $A_1(\lambda)$. Therefore, the bound of H will be dominated by the bound of $A_1(\lambda)$. First, let us get the bound for $\|A_1(\lambda)\|$. Recall the equation (38) that $G_\lambda^{-1}(\boldsymbol{f}) = A^{-1} + A^{-1}P(\mathbf{1} - P^TA^{-1}P)^{-1}P^TA^{-1}$. So,

$$\begin{aligned}\|G_\lambda^{-1}(\boldsymbol{f})\lambda W\| &= \|(I + A^{-1}P(\mathbf{1} - P^TA^{-1}P)^{-1}P^T)\lambda A^{-1}W\| \\ &\leq \|I + A^{-1}P(\mathbf{1} - P^TA^{-1}P)^{-1}P^T\| \cdot \|\lambda A^{-1}W\|\end{aligned}$$

Note the definition of G_1, where $G_1 = \mathrm{diag}(p_1 - p_1^2 + \lambda w_1, \ldots, p_{K-1} - p_{K-1}^2 + \lambda w_{K-1})$. Obviously, $\|A^{-1}\lambda W\|_2^2 = \|A^{-1}G_1G_1^{-1}\lambda W\|_2^2 \leq \|A^{-1}G_1\|_2^2\|G_1^{-1}\lambda W\|_2^2$. From Lemma 1, it is easy to see that $\|A^{-1}G_1\|_2 = \max_i\{\|(p_i + \lambda w_i)^{-1}(p_i + \lambda w_i - p_i^2)\|_2\} \leq \max_i\{1 + \|p_i\|_2\}$ Therefore, $\|A^{-1}\lambda W\|_2^2$ can be bounded by $\|G_1^{-1}\lambda W\|_2^2$. Using Lemma 7 and similar derivation as the proof of Proposition 1, it is clear that $\|G_1^{-1}\lambda W\|_2^2 = O(\lambda K)$. So we can see that the order for $\|\lambda A^{-1}W\|_2^2$ is $O(\lambda K)$. Furthermore,

$$\|A^{-1}P(\mathbf{1} - P^TA^{-1}P)^{-1}P^T\|_2 \leq \frac{\|A^{-1}P\|\|P\|}{1 - \sum_{i=1}^{K-1}\|p_i(p_i + \lambda w)^{-1}p_i\|}. \tag{115}$$

Using Lemma 2 and Lemma 8, it is clear that $\|(p_i + \lambda w_i)^{-1}\|_2^2 \leq \|(p_i - p_i^2 + \lambda w_i)^{-1}\|_2^2 \leq O(K^2(c_\lambda K)^{-\frac{1}{2m}})$, where c_λ is a constant independent of K. Then we can obtain

$$\|A^{-1}P\| \leq \sum_{i=1}^{K-1}\|(p_i + \lambda w_i)^{-1}p_i\|_2^2 \leq \sum_{i=1}^{K-1}[\frac{1}{K^2}O(K^2(c_\lambda K)^{-\frac{1}{2m}})] = O(K(c_\lambda K)^{-\frac{1}{2m}}); \tag{116}$$

$$\|P\|_2^2 = \sum_{i=1}^{K-1}\|p_i\|_2^2 \leq \sum_{i=1}^{K-1}\frac{1}{K^2} = O(\frac{1}{K}). \tag{117}$$

Therefore,

$$\begin{aligned}\|A^{-1}P(\mathbf{1}-P^TA^{-1}P)^{-1}P^T\|_2 &\leq \frac{\sqrt{O(K(c_\lambda K)^{-\frac{1}{2m}})}\sqrt{O(\frac{1}{K})}}{1-\sum_{i=1}^{K-1}\sqrt{O\left(\frac{1}{K^4}K^2(c_\lambda K)^{-\frac{1}{2m}}\right)}} \\ &= \left|\frac{\sqrt{O((c_\lambda K)^{-\frac{1}{2m}})}}{1-\frac{K-1}{K}\sqrt{O((c_\lambda K)^{-\frac{1}{2m}})}}\right| \\ &= O(1). \end{aligned} \tag{118}$$

Hence, we get the conclusion that $\|(I + A^{-1}P(\mathbf{1} - P^TA^{-1}P)^{-1}P^T)\| = O(1)$, and $\|A_1(\lambda)\| = \|G_\lambda^{-1}(f)\lambda W\|_2^2 = O(\lambda K)$. Based on the definition of H, we obtain that

$$\|H\|_2^2 \leq O(\|A_1(\lambda)\|_2^2) = O(\lambda K). \tag{119}$$

□

Lemma 13. *Under the conditions in Theorem 3, and let Λ denote $K \times K$ diagonal matrix with diagonal entries $\Lambda_{11} > \Lambda_{22} > \cdots > \Lambda_{L_0L_0} = \cdots = \Lambda_{KK} = 0$. Assume that there exists a D^* such that $|\Lambda_{rr} - \Lambda_{ss}| > D^*\tilde{\lambda}_r$ for $r \neq s$.*

Let Ξ denote a symmetric $K \times K$ random matrix with entries ξ_{rs}, $r, s = 1, \ldots, K$. Assume that there exists a sequence δ_K with the following properties: (1) $K\delta_K \to 0$ as $K \to \infty$; (2) $\sup_{r,s}(E\xi_{rs})^2/(\tilde{\lambda}_r\tilde{\lambda}_s)) = O(\delta_K)$ as $K \to \infty$. The the following hold:

$$|\lambda_r(\Lambda + \Xi) - \Lambda_{rr}| = O\left((E\xi_{rr}^2)^{1/2} + \sum_{s=1}^{K}\frac{E\xi_{rs}^2}{\max\{\tilde{\lambda}_r, \tilde{\lambda}_s\}}\right),\ r = 1, \ldots K; \tag{120}$$

and

$$\|\gamma_r(\Lambda + \Xi) - \gamma_r(\Lambda)\|_2^2 = O\left(\frac{\sum_{s\neq r}\left[E\xi_{rs}^2/\max\{\tilde{\lambda}_r, \tilde{\lambda}_s\}\right]}{\tilde{\lambda}_r}\right),\ r = 1, \ldots K. \tag{121}$$

Lemma 14. *Notation follows Section 4.3.3. Suppose ξ_{rs} is the rs'th element of matrix Ξ, where Ξ is defined as*

$$\Xi = \frac{1}{n}Q^T[\tilde{F}(B \cdot D\tilde{F}) + (B \cdot D\tilde{F})^T\tilde{F}^T + (B \cdot D\tilde{F})^T(B \cdot D\tilde{F})]Q. \tag{122}$$

Here $B = -\left(I + G_\lambda^{-1}(\tilde{\boldsymbol{f}})[D^2 l_n(\tilde{\boldsymbol{f}}) - D^2 l(\tilde{\boldsymbol{f}})]\right)^{-1} G_\lambda^{-1}(\tilde{\boldsymbol{f}})$ *and* $D\tilde{F} = (d\tilde{\boldsymbol{f}}(x_1), \ldots, d\tilde{\boldsymbol{f}}(x_n))^T$. *Then we can obtain*

$$E\xi_{rs}^2 = O(K\frac{K}{n}(\lambda K)^{-\frac{1}{2m}}). \tag{123}$$

Proof: Recall the construction of Q. Suppose that the matrix Q in (122) is $Q = (\mathbf{q}_1, \ldots, \mathbf{q}_L)$, then ξ_{rs} can be written as

$$\frac{1}{n}\mathbf{q}_r^T[\tilde{F}(B \cdot D\tilde{F}) + (B \cdot D\tilde{F})^T\tilde{F}^T + (B \cdot D\tilde{F})^T(B \cdot D\tilde{F})]\mathbf{q}_r \tag{124}$$

By ignoring the smaller order terms, it is easy to see that the order of $E\xi_{rs}^2$ can be bounded by $E\|\frac{1}{n}\mathbf{q}_r^T[\tilde{F}(B \cdot D\tilde{F})]\mathbf{q}_s\|^2$ up to some constant, i.e.,

$$E\xi_{rs}^2 = O(E\|\frac{1}{n}\mathbf{q}_r^T[\tilde{F}(B \cdot D\tilde{F})]\mathbf{q}_s\|^2). \tag{125}$$

Now we focus on getting an upper bound for (125). Note that $Q^TQ = I$, which implies that both $\mathbf{q}_r$ and $\mathbf{q}_s$ are normalized vectors. Recall the definition of $\tilde{F}$ and $D\tilde{F}$, where $\tilde{F} = (\tilde{\boldsymbol{f}}(x_1), \ldots, \tilde{\boldsymbol{f}}(x_n))^T$and $D\tilde{F} = (d\tilde{\boldsymbol{f}}(x_1), \ldots, d\tilde{\boldsymbol{f}}(x_n))^T$. Using the fact $\|A\|_2 \leq \sqrt{mn}\max_{ij}|a_{ij}|$ for any matrix $A \in \mathbb{R}^{m\times n}$, we can obtain that

$$\begin{aligned}\|\frac{1}{n}\mathbf{q}_r^T[\tilde{F}(B \cdot D\tilde{F})]\mathbf{q}_s\|_2^2 &\leq \frac{1}{n^2}\|\tilde{F}(B \cdot D\tilde{F})\|_2^2 \\ &\leq (\max_{i,j}|\tilde{\boldsymbol{f}}^T(x_i)Bd\tilde{\boldsymbol{f}}(x_j)|^2).\end{aligned} \tag{126}$$

Here both $\tilde{\boldsymbol{f}}$ and $d\tilde{\boldsymbol{f}}$ are vectors, and $B = -\left(I + G_\lambda^{-1}(\tilde{\boldsymbol{f}})[D^2 l_n(\tilde{\boldsymbol{f}}) - D^2 l(\tilde{\boldsymbol{f}})]\right)^{-1} G_\lambda^{-1}(\tilde{\boldsymbol{f}})$, which is an operator defined in Section 4.3.3. From the proof of Proposition 3, we can easily know that the order of the operator $G_\lambda^{-1}(\tilde{\boldsymbol{f}})[D^2 l_n(\tilde{\boldsymbol{f}}) - D^2 l(\tilde{\boldsymbol{f}})$ is $o_K(1)$ as the number of data points $n \to \infty$. Hence the operator $I + G_\lambda^{-1}(\tilde{\boldsymbol{f}})[D^2 l_n(\tilde{\boldsymbol{f}}) - D^2 l(\tilde{\boldsymbol{f}})]$ will be dominated by I up to some terms with smaller order. Furthermore, the operator B can be dominated by $G_\lambda^{-1}(\tilde{\boldsymbol{f}})$, of which we have known some properties. We also obtained that $\tilde{\boldsymbol{f}} - \boldsymbol{f}^0 = -G_\lambda^{-1}(\tilde{\boldsymbol{f}})Z_\lambda(\boldsymbol{f}^0)$ from the first part. Therefore, we can derive

the following

$$|\tilde{\boldsymbol{f}}(x_i)Bd\boldsymbol{f}(x_j)| \leq O(|[\boldsymbol{f}^0 - G_\lambda^{-1}(\tilde{\boldsymbol{f}})Z_\lambda(\boldsymbol{f}^0)]^T G_\lambda^{-1}(\tilde{\boldsymbol{f}})df(x_j)|)$$
$$\leq O(|(\boldsymbol{f}^0)^T(x_i)G_\lambda^{-1}(\tilde{\boldsymbol{f}})d\boldsymbol{f}(x_j)|). \tag{127}$$

The last inequality is provided by ignoring the smaller order terms.

From the proof of Lemma 12 , it is clear that the operator $G_\lambda^{-1}(\tilde{\boldsymbol{f}})$ can be bounded by G_1^{-1}, where $G_1 = \mathrm{diag}(p_1 - p_1^2 + \lambda w_1, \ldots, p_{K-1} - p_{K-1}^2 + \lambda w_{K-1})$. Then from (127), we have that

$$|(\boldsymbol{f}^0)^T(x_i)G_\lambda^{-1}(\tilde{\boldsymbol{f}})d\boldsymbol{f}(x_j)|^2 = |\sum_{k=1}^{K-1} f_k^0(x_i)(p_k - p_k^2 + \lambda w_k)^{-1}df_k(x_j)|^2$$
$$\leq C\sum_{k=1}^{K-1} |f_k^0(x_i)(p_k - p_k^2 + \lambda w_k)^{-1}df_k(x_j)|^2, \tag{128}$$

where C is a constant. From Lemma 8, it is known that $\|(p_k - p_k^2 + \lambda w_k)^{-1}df_k(x_j)\|_2^2 = O(\frac{K}{n}(\lambda K)^{-\frac{1}{2m}})$, then

$$|(\boldsymbol{f}^0)^T(x_i)G_\lambda^{-1}(\tilde{\boldsymbol{f}})df(x_j)|^2 \leq O(\frac{K}{n}(\lambda K)^{-\frac{1}{2m}})\sum_{k=1}^{K-1}(f_k^0(x))^2. \tag{129}$$

Next, we come to show that $\sum_{k=1}^{K-1}(f_k^0(x))^2$ is bounded by the order K. Recall that the multi-logit model is

$$p_k(\boldsymbol{f}(x)) \triangleq P(Y = k|X) = \frac{e^{f_k(x)}}{1 + \sum_{j=1}^{K-1} e^{f_j(x)}},$$

where $X \in \mathbb{R}^d$. The function vector f is coded with zero constraint, i.e., $\boldsymbol{f} = (f_1, \ldots, f_{K-1}, 0)$.

Note that $p_k(\boldsymbol{f}(x))$ is bounded away from zero and one. We also have assumed that $p_i \sim 1/K$, $i = 1, \ldots, K$, then there exists a constant C_2 independent of number of classes K, such that

$$\sum_{i=1}^{K-1}(\log\frac{p_i}{p_K})^2 \leq C_2 K, \tag{130}$$

The derivation is as following. Using the Taylor expansion, we know that

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$$

By ignoring the high order term, we can have

$$\log \frac{p_i}{p_K} = \log(1 + \frac{p_i}{p_K} - 1) \approx \frac{p_i}{p_K} - 1.$$

Hence there exists a constant C_2 satisfying

$$\sum_{i=1}^{K-1} (\log \frac{p_i}{p_K})^2 \approx \sum_{i=1}^{K-1} \left(1 - \frac{p_i}{p_K}\right)^2 \leq C_2 K, \tag{131}$$

where the last inequality is provided by the assumption that each $p_i \sim 1/K$. Now we can get the bound $\sum_{k=1}^{K-1} (f_k^0(x))^2 \leq \sum_{i=1}^{K-1} (\log \frac{p_i}{p_K})^2 \leq O(K)$. Plugging it into (129), we finish the proof. $\square$

REFERENCES

Albert, A. and Anderson, J. A. (1984). On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, **71**, 1–10.

Allwein, E. L., Schapire, R. E. and Singer, Y. (2000). Reducing multiclass to binary: A unifying approach for margin classifiers, *Proc. 17th International Conf. on Machine Learning* , Morgan Kaufmann, San Francisco, CA, 9–16.

Bai, X. D., Gao, P. X., Wang, Z. L., and Wang, E. G. (2003). Dual-mode Mechanical Resonance of Individual ZnO Nanobelts. *Applied Physics Letters*, **82**, 4806–4808.

Benham, P.P. and Crawford , R.J. (1987). *Mechanics of Engineering Materials*, John Wiley&Sons, New York.

Boser, B., Guyon, I., and Vapnik, V. (1992). A Training Algorithm for Optimal Margin Classifiers. In *Proceedings Fifth ACM Workshop on Computational Learning Theory*, pp. 144–152.

Bregler, C. and Omohundro, M. (1994). Surface Learning with Applications to Lipreading. In J. D. Cowan, G. Tesauro, and J. Alspector (eds.), *Advances in Neural Information Precessing Systems 6*, Morgan Kaufmann, San Francisco.

Breiman, L. (1995). Better Subset Regression using the Nonnegative Garrote. *Technometrics*, **37**, 373-384.

Bickel, P. J. and Levina, E. (2007). Regularized Estimation of Large Covariance Matrices. *Ann. Statist.*, To appear.

Chen, C.Q., et al. (2006). Size Dependence of Young's Modulus in ZnO Nanowires. *Physical Review Letters*, **96**, 075505.

Campbell, C., Cristianini, N. and Smola, A. (2000). Query Learning with Large Margin Classifiers. In *Proceedings of 17th International Conference on Machine Learning*, pages 111–118.

Choulakian, V. (2005). Transposition Invariant Principal Component Analysis in L1 for Long Tail Data. *Statistics and Probability Letters*, **71**, 23-31.

Cohn, D. A., Ghahramani, Z. and Jordan, M. I. (1996). Active Learning with Statistical Models. *Journal of Artificial Intelligence Research*, **4**, 129–145.

Cook, R. and Weisberg, S. (1984). *Residuals and Influence in Regression*, Chapman and Hall, London.

Cox, D. and O'Sullivan, F. (1990). Asymptotic Analysis of Penalized Likelihood and Related Estimators. *Annals of Statistics*, **18**, 1676–1695.

Critchley, F. (1985). Influence in Principal Components Analysis. *Biometrika*, **72**, 627-636.

Croux, C. and Haesbroeck, G. (2000). Principal Component Analysis Based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies. *Biometrika*, **87**, 603–618.

Dempster, A. (1972). Covariance Selection. *Biometrics*, **28**, 157–75.

Dey, D. K. and Srinivasan, C. (1985). Estimation of a Covariance Matrix under Steins Loss. *Ann. Statist.*, **13(4)**, 1581–1591.

Dietterich, T. G. and Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes, *Journal of Artificial Intelligence Research* **2**: 263–286.

Duda, R. O., Hart, P. E. and Stork, D. G. (2000), *Pattern Classification*, New York: John Wiley & Sons.

Fedorov, V. V. (1972), *Theory of Optimal Experiments*, Academic Press, New York.

Fukumizu, K. (2000), Statistical Active Learning in Multilayer Perceptrons, *IEEE Transactions on Neural Networks*, 11(1), 17–26.

Haff, L. R. (1980). Empirical Bayes Estimation of the Multivariate Normal Covariance Matrix. *Ann. Statist.*, 8(3):586–597.

Hastie, T. and Stuetzle, W. (1989). Principal Curve. *Journal of The American Statistical Association*, **84**, 502-516.

Hastie, T. and Tibshirani, R.(1990) *Generalized Additive Models.* London, Chapman and Hall.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance Matrix Selection and Estimation via Penalised Normal Likelihood. *Biometrika*, **93(1)**, 85–98.

Ibazizen M. and Dauxois J. (2003). A Robust Principal Component Analysis. *Statistics*, **37**, 73–83.

Jackson, J. (1991). *A Users Guide to Principal Components*, Wiley, New York.

Joseph, V. R. (2004). Efficient Robbins-Monro Procedure for Binary Data. *Biometrika*, **91**, 461–470.

Joseph, V.R. and Melkote, S.N. (2009). Statistical Adjustments to Engineering Models. *Journal of Quality Technology*, to appear.

Joseph, V. R., Tian, Y. and Wu, C. F. J. (2007). Adaptive Designs for Stochastic Root-Finding. *Statistica Sinica*, **17**, 1549–1565.

Jolliffe, I. (1986) *Principal Component Analysis*, Springer-Verlag, New York.

Kato, T. (1996). *Perturbation Theory for Linear Operators*. Springer, New York.

Kiefer, J. (1959). Optimum Experimental Designs. *Journal of the Royal Statistical Society, Series B*, **21**, 272–304.

Kneip A. (1994). Nonparametric Estimation of Common Regressors for Simliar Curve Data, *The Annals of Statistics*, **22**, 1386-1427.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1990). Handwritten Digit Recognition with a Back-Propagation Network. *in* Touretzky, D. (ed.), *Advances in Neural Information Processing Systems*, Vol. 2, Morgan Kaufman, Denver, CO.

Ledoit, O. and Wolf, M. (2003). A Well-conditioned Estimator for Large-dimensional Covariance Matrices. *Journal of Multivariate Analysis*, **88**, 365–411.

Lee Y., Lin, Y. and Wahba, G. (2004). Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *J. Am. Statist. Assoc.* **99**, 67–81.

Levina, E., Rothman, A. J., and Zhu, J. (2007). Sparse Estimation of Large Covariance Matrices via a Nested Lasso Penalty. *Annals of Applied Statistics*, To appear.

Lewis, D. and Gale, W. (1994). A Sequential Algorithm for Training Text Classifiers. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, Springer-Verlag.

Lin, Y. (2000). Tensor Product Space ANOVA Models. *Annals of Statistics*, **28**, 734–755.

Lin, Y., Lee, Y. and Wahba, G. (2002). Support Vector Machines for Classification in Nonstandard Situations. *Machine Learning*, **46**, 191–202.

MacKay, D. J. C. (1992). Information-Based Objective Functions for Active Data Selection. *Neural Computation*, **4(4)**, 590–604.

Mai, W. J. and Wang, Z.L. (2006). Quantifying the Elastic Deformation Behavior of Bridged Nanobelts. *Applied Physics Letters*, 073112.

McLeish, D. L. and Tosh, D. (1990). Sequential Designs in Bioassay. *Biometrics*, **46**, 103–116.

Miller, A. (2002). *Subset Selection in Regression*. CRC Press, New York.

Neyer, B. T. (1994). D-Optimality-Based Sensitivity Test. *Technometrics*, **36**, 61–70.

Oja, E. (1982). A Simplified Neuron Model as a Principal Component Analyzer. *Journal of Mathematical Biology*, **15** , 267–273.

Oja, E. , Ogawa, H. and Wangviwattana, J. (1991). Learning in Nonlinear Constrained Hebbian Networks. In *Artificial Neural Networks*, Elsevier, Amsterdam, pp. 385 – 390.

Pan, Z.W., Dai, Z.R. and Wang, Z.L. (2001). Nanobelts of Semiconducting Oxides. *Science* **291**, 1947-1949.

Perron, F. (1992). Minimax Estimators of a Covariance Matrix. *Journal of Multivariate Analysis*, **43**, 16–28.

Poncharal, P., Wang, Z.L., Ugarte, D. and de Heer, W.A. (1999). Electrostatic Deflections and Electromechanical Resonances of Carbon Nanotubes. *Science* **83**, 1513–1516.

Pourahmadi, M. (1999). Joint Mean-Covariance Models with Applications to Longitudinal Data: Unconstrained Parameterisation. *Biometrika*, **86**, 677–690.

Pourahmadi, M. (2000). Maximum Likelihood Estimation of Generalized Linear Models for Multivariate Normal Covariance Matrix. *Biometrika*, **87**, 425–435.

Pukelsheim, F. (1993), *Optimal Design of Experiments*, New York: John Wiley & Sons.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, **22**, 400–407.

Robert E. Schapire (1997). Using output codes to boost multiclass learning problems. *In Machine Learning: Proceedings of the Fourteenth International Conference*, pages 313–321.

San Paulo, A., et al. (2005). Mechanical Elasticity of Single and Double Clamped Silicon Nanobeams Fabricated by the Vapor-liquid-solid Method. *Applied Physics Letters*, 053111.

Santner, T. J. and Duffy, D. E. (1986). A Note on A. Albert and J. A. Andersons Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, **73**, 755–758.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer, New York.

Salvetat et al. (1999). Elastic and Shear Moduli of Single-Walled Carbon Nanotube Ropes. Physical Review Letters, **82**(5), 944–947.

Salvetat et al. (1999). Mechanical Properties of Carbon Nanotubes. *Applied Physics A*, **69**, 255–260.

Schohn, G. and Cohn, D. (2000). Less is More: Active Learning with Support Vector Machines. In *Proceedings of the Seventeenth International Conference on Machine Learning*.

Schölkopf, B., Smola, A. and Müller, K. (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Computation*, (5) **10**, 1299–1319.

Schölkopf, B. and Smola, A. (2002). *Learning with Kernels*, MIT Press, Cambridge.

Silvapulle, M. J. (1981). On the Existence of Maximum Likelihood Estimators of the Binomial Response Model. *Journal of the Royal Statistical Society, Series B*, **43**, 310–313.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, **10**, 795–810.

Silvey, S. D. (1980), *Optimal Design*. London: Chapman and Hall.

Smith, M. and Kohn, R. (2002). Parsimonious Covariance Matrix Estimation for Longitudinal Data. *J. Amer. Statist. Assoc.*, **97(460)**, 1141–1153.

Song, J. H., Wang, X. D., Riedo, E., and Wang, Z.L. (2005). Elastic Property of Vertically Aligned Nanowires. *Nano Letters*, **5**, 1954.

Stein, C. (1977). Lectures on the Theory of Estimation of Many Parameters. In *Studies in the Statistical Theory of Estimation, Part I* (I. A. Ibragrniov and M. S. Nikulin, eds.). *Proc. Scientific Seminars Steklov Institute, Leningrad Division*, **74**, 4–65. (In Russian.)

Stone, C.J. (1985) Additive regression and other nonparametric models, *The Annals of Statistics*, **13**, 689–705.

Tewari, A. and Bartlett, P .L. (2007). On the consistency of multiclass classification methods, *Journal of Machine Learning Research*, **8**, 1007-1025.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, **58**, 267–288.

Tong, S. and Koller, D. (2001). Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research*, **2**, 45–66.

Tütüncü, R. H., Toh, K. C., and Todd, M. J. (2003). Solving Semidefinite-quadratic-linear Programs Using SDPT3. *Mathematical Programming*, *95(2)*, 189–217.

Vapnik, V. and Chervonenkis, A. (1974) *Theory of Pattern Recognition*, Nauka, Moscow.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM, Philadephia.

Wang, Z.L. and Song, J.H. (2006). Piezoelectric Nanogenerators Based on Zinc Oxide Nanowire Arrays. *Science*, *312*, 242-246.

Winkler, W. E. (2006). *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*, New York: Springer.

Wong, E. W., Sheehan, P. E., Lieber, C. M. (1997). Elasticity, Strength and Toughness of Nanorods and Nanotubes. *Science* **277**, 1971-1975.

Wong, F., Carter, C., and Kohn, R. (2003). Efficient Estimation of Covariance Selection Models. *Biometrika*, **90**, 809–830.

Wu, B., Heidelberg, A., and Boland, J.J. (2005). Mechanical Properties of Ultrahigh-Strength Gold Nanowires. *Nature Materials*, **4**, 525-529.

Wu, C. F. J. (1985). Efficient Sequential Designs with Binary Data. *Journal of the American Statistical Association*, **80**, 974–984.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric Estimation of Large Covariance Matrices of Longitudinal Data. *Biometrika*, **90**, 831–844.

Ying, Z. and Wu, C. F. J. (1997). An Asymptotic Theory of Sequential Designs Based on Maximum Likelihood Recursions. *Statistica Sinica*, **7**, 75–91.

Young, L. J. and Easterling, R. G. (1994). Estimation of Extreme Quantiles Based on Sensitivity Tests: A Comparative Study. *Technometrics*, **36**, 48–60.

Yu, M. F., et al. (2000). Strength and Breaking Mechanism of Multiwalled Carbon Nanotubes under Tensile Load. *Science* **287**, 637–640.

Yuan, M. and Lin, Y. (2007). Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*, **94(1)**, 19–35.

Zhang T. (2004). Statistical Analysis of Some Multi-Category Large Margin Classification Methods, *Journal of Machine Learning Research*, **5**, 1225-1251.

Zhou, J. et al. (2006). Nanowire as Pico-gram Balance at Workplace Atmosphere. *Solid State Communications*, **139**, 222–226.